

WG/GAML/11/2.4

ESTIMATING BENCHMARKS FOR FOUNDATIONAL SKILLS IN RELATION TO TARGETS FOR SDG 4.1.1: THE CASE OF READING



FEBRUARY 2025





Contents

Executive Summary.....	1
Section 1: Introduction	2
Section 2: Background to this Study	3
Section 3: Theoretical Framework and Initial Benchmark Analysis Results	5
Section 4: Updated Results from TAG Recommendations.....	31
Section 5: Data Analysis of Additional Countries and Languages Using Accuracy-Based Model	44
Section 6: Conclusions and Summary.....	50
References	54

DRAFT

Executive Summary¹

This study aims to establish a more rigorous psychometric framework for analyzing foundational literacy assessments, which are typically oral, one-on-one, and conceptually complex. These assessments have lacked the comparability and benchmarking capabilities of larger international and regional assessments. The study's findings demonstrate that Item Response Theory (IRT), particularly the Rasch model, is an effective method for setting benchmarks in literacy assessments, especially for reading comprehension (RC) and oral reading fluency (ORF). Unlike Generalized Linear Mixed Models (GLMM), which are computationally efficient but fail to account for item-level differences, IRT provides a unified latent scale for comprehension and precursor skills, offering deeper insights into student ability.

However, the study highlights trade-offs between benchmarking approaches. Traditional methods offer precise but subjective benchmarks, GLMM sets benchmarks with high statistical precision but without item-level context, and IRT provides moderate precision while incorporating the most comprehensive data available. Although IRT confidence intervals may be broader than ideal from a reading science perspective, they remain acceptable from a psychometric standpoint.

The analysis also reinforces the unidimensionality and reliability of foundational reading assessments, confirming that certain skills (e.g., accuracy and ORF) are stronger predictors of reading proficiency than others (e.g., letter sounds and listening comprehension). Additionally, the study represents one of the largest compilations of foundational reading data to date, spanning 32 languages across eight countries. This dataset offers unique insights into language group trends, though further modeling is needed for broader linguistic comparisons.

While IRT provides a theoretically sound approach to benchmark setting, its limitations include broad confidence intervals for non-RC subtasks and the assumption of unidimensionality. Future research should explore alternative IRT models (2PL, 3PL, and multidimensional IRT), Bayesian estimation methods for small sample sizes, and refinements in confidence interval estimation. Despite these challenges, IRT remains the most comprehensive method for setting literacy benchmarks, leveraging item-level data to enable meaningful comparisons across assessments and educational contexts.

¹ Authored by Abdullah Ferdous and Eric Muller - American Institutes for Research.

Section 1: Introduction

The United Nations (UN) Sustainable Development Goal (SDG) 4 aims to ensure that, by 2030, “all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.” SDG 4.1.1 measures the proportion of children and young people achieving at least a minimum proficiency level (MPL) in reading and math at three educational stages: Grades 2/3 (early primary), end of primary, and end of lower secondary.


The UNESCO Institute for Statistics (UIS) is the custodian agency responsible for reporting progress on education SDG, and develops standards, methodologies and guidelines to enable countries to report on these goals. Since 2016, UIS has worked through the [Global Alliance to Monitor Learning \(GAML\)](#) to support national strategies for learning assessments and developing internationally comparable indicators and methodological tools to measure progress towards SDG 4 targets.

However, reporting on SDG 4.1.1a (Grades 2/3 proficiency in reading and mathematics) has been insufficient:

- As of late 2023, only 37 countries reported learning data at the Grade 2/3 level.
- Over the past six years, only 101 countries reported learning data at the end-of-primary, while 203 countries reported primary school enrollment.

The slow increase in reporting rates of 4.1.1a in the past several years suggest it could take decades to match enrollment reporting rates. Several challenges contribute to the lack of coverage for 4.1.1a. First, many countries use children’s first languages for instruction in the early grades. Different orthographies affect reading acquisition in complex ways, potentially requiring benchmarks specific to different language groups. Second, many widely used early grade learning assessments were designed for advocacy and monitoring rather than international proficiency reporting. Finally, there are more significant technical difficulties in measuring skills at the lower primary level when children are in the formative phases of learning to read. There is a risk of evaluating a conditioned response rather than cognitive skill.

Due to low reporting rates, indicator 4.1.1a was downgraded from a Tier 1 indicator to a Tier 2 indicator in October 2023 by the [UN’s Interagency Expert Group on the SDGs](#) (IAEG-SDGs). The reclassification signifies that the indicator lacks adequate data for meaningful cross-national comparisons. Indicator 4.1.1 parts b (end of primary) and c (end of lower secondary) remain Tier 1 indicators. As a Tier 2 indicator, 4.1.1a risks removal from the SDG framework, as the IAEG-SDGs plans to eliminate all Tier 2 indicators in 2025. To restore SDG 4.1.1a as a Tier 1 indicator, at least 50% of countries where the indicator is relevant must report data.



Global education leaders and the IAEG agreed on this goal and have set out action steps to provide better guidance and information to countries about available tools for global reporting. This report documents the results of a study undertaken as a key step to improve reporting: developing a robust benchmarking methodology for reporting on indicator 4.1.1a reading skills.


Section 2: Background to this Study

In December 2023, UIS, GAML experts, and key stakeholders convened to discuss strategies for increasing SDG 4.1.1a coverage while maintaining rigorous methodological standards. The discussions focused on resolving challenges in producing reliable and internationally comparable learning outcomes data, reviewing recent advancements, and outlining an agenda for future improvements.

One of the main concerns addressed was the limited use of well-known and commonly administered assessments for 4.1.1a reporting. These include the Early Grade Reading Assessment (EGRA), the Foundational Learning Module (FLM) of the Multiple Indicators Cluster Survey (MICS), and the People's Action for Learning (PAL) Network tools. These assessments were primarily designed for advocacy, program design, and monitoring and evaluation, not for global reporting and comparison. UIS has not accepted data from these assessments for 4.1.1a reporting due to the lack of explicit alignment with the Minimum Proficiency Levels (MPL) and Global Proficiency Framework (GPF), and insufficient documentation of their properties.

To address these issues, UIS was requested to prepare eligibility criteria for reporting against indicator 4.1.1.a, including both psychometric and procedural standards for assessments. The resulting [draft criteria](#) cover: 1) alignment to the MPL and construct validity, 2) item content and quality, 3) population coverage and sampling, 4) assessment administration and data custodianship, 5) reliability, 6) benchmark-based linking to the MPL, and 7) maintaining standards over time. UIS solicited input from the community of interest and a Technical Advisory Group (TAG) tasked with commenting and advising on the criteria and issues related to measurement.

Based on revisions and recommendations to the criteria from the first TAG meeting in March 2024, UIS proposed to the IAEG-SDGs to unpack SDG 4.1.1a reporting to address two key measurement issues. First, home languages used for instruction in early grades pose measurement and benchmarking challenges due to differences in orthographic complexity. Second, early grade children are still mastering precursor elements of reading (e.g., oral language comprehension and decoding with fluency), complicating traditional assessments. The MPL for SDG 4.1.1 focuses on the transition from 'learning to read' to 'reading to learn'



(i.e. reading comprehension).² In the early grades, measuring precursor skills involves measuring a conditioned response rather than cognition (e.g., one is trying to assess automaticity in decoding, which is different from the cognition required for processing comprehension questions). This presents challenges to conventional psychometric techniques.

UIS therefore proposed benchmarks for reading precursor skills, allowing countries to measure and report progress towards the MPL in the early grades. As many assessments such as EGRA, MICS, and PAL Network address precursor skills, AIR used their data for a benchmark analysis.

Results of the benchmark analysis (presented below in Section 3) were discussed during the second TAG meeting in May 2024. The TAG concluded that high-quality data and large sample sizes could, in principle, support the establishment of language-specific benchmarks, drawing on insights from South Africa and Kenya’s national methodologies. However, further analysis was required to assess method reliability, and additional data was needed for certain languages as well as for numeracy and mathematics—leading to the additional results discussed in Sections 4 and 5 below). The TAG also highlighted the importance of clearer definitions of assessment difficulty levels and expanded national benchmark-setting experiences. UIS is currently addressing these issues through a separate initiative, distinct from the scope of this study.

The TAG meetings in March and May confirmed that all assessments meeting the eligibility criteria can be used for reporting, including traditional assessments, newer assessments that measure precursor skills, and national assessments. The recommended metrics for reporting are:

- **Reading:** The percentage of children correctly answering a sufficient number of comprehension questions.
- **Mathematics:** Initially, the percentage of children correctly answering questions on numbers and operations.

This focus on numbers and operations, rather than other areas of mathematics (e.g. shapes, early algebraic skills), is a practical choice. While all areas are equally important for achieving the MPL in mathematics, the TAG recommended starting with numbers and operations while research on benchmarks for other mathematics components is conducted.

Countries with a low percentage of children demonstrating reading with understanding or strong mathematical skills may report on precursor skills in reading, and broader mathematical abilities beyond operations. These data can be compared to benchmarks to

² UIS has come to refer to all skills “prior” to comprehension as “precursor” skills. This may not be conventional terminology in reading science but is a convenient shorthand.

track. However, to be useful, benchmarks must have numerical values and should be established using a defensible methodology. The remainder of this study report outlines a method for developing benchmarks in reading and provides examples of its application. Benchmarks for mathematics are under development.


Section 3: Theoretical Framework and Initial Benchmark Analysis Results

In the hierarchy of foundational literacy skills, Reading Comprehension (RC) represents the apex, built upon several precursor skills, including Oral Reading Fluency (ORF), Listening Comprehension (LC), Letter Sound (LS), Syllable Sound (SS), Familiar Words (FMW), Invented Words (INW), and Silent Reading Comprehension (SRC). Traditionally, educators and administrators have set benchmarks for subtasks involving multiple-choice or binary response items, such as RC, by adopting thresholds—typically 60% or 80% correct responses—which have become customary in the communities of practice working with precursor skills assessment (e.g., [EGRA Benchmarks](#); [EGRA Benchmarking Process](#)). These thresholds align with what many school systems consider “passing” (60% to 69%) and “good” (80% to 89%) ([Academic Grading in the United States](#)) and were endorsed as reasonable benchmarks during the TAG meeting in March 2024.

The TAG recommended reporting only RC proficiency but emphasized the importance of supporting low- and middle-income (LMI) countries by also reporting on the percentage of learners who have mastered precursor skills. This approach provides critical insights into students’ progress along the learning pathway toward reading comprehension benchmarks, allowing countries to make informed decisions about interventions that improve educational outcomes.

However, linking RC benchmarks to benchmarks in the precursor skills presents several challenges. Some precursor assessments involve timed tasks, while others are timed though not explicitly time-limited. For instance, passage-based ORF, or letter-sounding, typically involve students reading a passage of reasonable word length or a list of letters within a one-minute timeframe. In these cases, students are assessed based on the number of words or letters read correctly per unit of time.

Although benchmarks exist for some precursor skills, they are only established in a limited number of countries. Additionally, these benchmarks are often norm-based (e.g., DIBELS) rather than criterion-referenced (e.g., EGRA in Lebanon Morocco), making it difficult to directly relate precursor skill performance to desired comprehension levels. While research suggests that precursor skills causally predict reading comprehension outcomes, translating a fixed percentage threshold (e.g., 80% or 60% for RC) into an ORF performance scale remains



ambiguous. This ambiguity complicates the process of setting meaningful proficiency standards for precursor skills.

To clarify these issues, we first examined two existing approaches to benchmark setting. Efforts to establish separate benchmarks for RC and precursor skills using **judgmental standard setting methods** have encountered several challenges:

- Cognitive burden on subject matter experts (SME). Experts must evaluate multiple precursor skills (typically six to seven), making the process complex and demanding.
- Inconsistencies in benchmarking across skills. The relationship between RC and precursor skill benchmarks varies across datasets, lacking clear coherence.
- Judgmental errors, particularly in shorter tasks. Tasks like listening comprehension are brief, raising questions about whether SMEs are genuinely consistent or if the consistency is forced due to the brevity of the task.

Despite these limitations, existing research supports the notion that foundational literacy skills form a unidimensional construct. This study confirms that finding and aligns with prior work in this area.

Previous benchmarking approaches that have attempted to establish **criterion-referenced proficiency standards** by linking precursor skills to RC, typically do not integrate this unidimensionality into joint benchmarking setting. Instead, they rely on classical statistical methods, assuming that comprehension questions and precursor skill tasks (e.g. recognizing words, letters, or passages) are equal in difficulty. These methods are not based on a theoretical framework that includes latent reading ability, making it difficult to accurately measure and align skill progression.

To address these gaps, this study proposed a **data-driven approach** for establishing RC and precursor skill benchmarks using Item Response Theory (IRT). This method allows for:

- More reliable benchmarking across multiple contexts (e.g. different countries, grade levels, and languages).
- A unifying measure of latent reading ability, maintaining the criterion-referenced approach of linking benchmarks to an agreed-upon RC proficiency level.
- Adjustments for fluency-based precursor skills, particularly those using timed measures that assess fluency across entire passages rather than specific words.

Rasch Accuracy IRT Model: Initial Approach

The first step in the proposed data-driven approach for setting benchmarks for RC precursor skills (i.e. timed subtasks such as ORF) by using pre-determined RC benchmarks (e.g. 60% or 80% accuracy) is IRT item calibration. IRT item calibration is a statistical process used to estimate the characteristics of assessment items within a framework that relates an individual's ability to their probability of responding correctly to an item (de Ayala, 2009).

According to IRT, each item on an assessment has numerical values for properties known as “parameters.” The parameters associated with an item depend on the chosen IRT model, but the most common parameters measured in IRT are:

- Difficulty – the ability level at which a respondent has a 50% chance of answering the item correctly.
- Discrimination – how well the item distinguishes between individuals with different ability levels.
- Guessing (in some models) – the probability of answering the item correctly by chance.

There are many IRT models, but for dichotomously scored data the three most common are (Rasch, 1960):

- Rasch Model (1 parameter) – only measures the difficulty parameter.
- 2PL (2 parameter) Model – measures the difficulty and discrimination parameter.
- 3PL (3 parameter) Model – measures the difficulty, discrimination, and a guessing parameter.

In this study, a Rasch model was used to analyze the data and obtain the item parameters. The Rasch model is designed to measure latent traits such as ability or proficiency and is characterized by its simplicity and strong mathematical properties, focusing on the relationship between a person's ability and the difficulty of the test items and putting these two concepts onto the same metric (Lord, 1980). The main assumption of the Rasch model is that the probability of a correct response depends only on the difference between the person's ability and the item's difficulty. Mathematically the Rasch model is expressed as:

$$P(X_{ji} = 1) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}$$


Where:

- $P(X_{ji} = 1)$ is the probability that person j answers item i correctly,
- θ_j is the ability level of person j ,
- b_i is the difficulty level of item i ,
- And e is the base of the natural logarithm (Embretson & Reise, 2000).

Key features of the Rasch model include (de Ayala, 2009; Lord, 1980):

- Unidimensionality: It assumes a single latent trait (e.g., ability) influences responses.
- Additivity: The model is based on the difference between ability level (θ_j) and difficulty level (b_i) making interpretation straightforward.
- Item Invariance: The difficulty of items is independent of the sample used for calibration.
- Person Invariance: A person's ability estimate is independent of the specific set of items administered.

The Rasch model is one of the most widely used IRT models in educational assessment due to its ease of interpretability, strong theoretical foundations, and ability to generate invariant



and comparable measurements across diverse groups and contexts. It, along with related models and tools in the same family, is employed in all major international and regional assessments that underpin SDG 4.1.1. Beyond its robust mathematical, statistical, and theoretical properties, its widespread adoption further justifies its selection for this report. The Rasch model is particularly valuable in educational assessment for calibrating items, setting benchmarks, and scaling scores.

Item Calibration

In IRT, the process of estimating item parameters based on item response data (i.e. responses collected from a sample's test items) is known as item calibration. Calibrating the Rasch model item parameters involves estimating each item's difficulty parameter (b_i) to create a scale that reflects the relationship between respondent ability (θ_j) and item difficulty (Baker & Kim, 2004). These parameter estimates are iteratively refined. In this study, estimation is conducted using the Expectation-Maximization (EM) algorithm, which maximizes the likelihood of the observed data under the model.

In the Rasch model, both item difficulty and person latent ability are represented on a continuous scale, ranging from negative infinity to positive infinity. In practice, however, most estimates fall within a narrower range, typically between -6 and 6, with values near zero representing average difficulty or ability. Higher positive values indicate more difficult items or higher ability, while lower negative values indicate easier items or lower ability.

For instance, an item with a difficulty parameter of 3.5 would be among the most challenging, requiring students in the highest percentiles to answer correctly. Conversely, an item with a difficulty of -3.5 would be relatively easy, with most students expected to answer correctly. Similarly, a respondent with ability estimates of 2.5 would be performing well above average, while a respondent with an estimate of -2.5 would be performing well below average. Because difficulty and ability estimates are often interpreted in relation to standard deviations, this study maintains the scale from -6 to 6 that aligns with this convention.

Item Calibration (via EM algorithm) Process:

1. **Data Collection** - Responses from a representative sample of assessment takers are gathered, with each response coded as correct (1) or incorrect (0).
2. **Initialization** - Initial estimates for item difficulties (b_i) and respondent abilities (θ_j) are generated, often using simple assumptions like uniform distributions.
3. **Expectation Step (E-Step)** - Using the current estimates of b_i and θ_j , the algorithm calculates the expected likelihood of the observed responses for each examinee-item interaction based on the Rasch model equation
4. **Maximization Step (M-Step)** - The item difficulty (b_i) and respondent ability (θ_j) estimates are updated to maximize the likelihood of the observed data.
5. **Iteration** - The E-step and M-step repeat until convergence, where changes in parameter estimates fall below a predefined threshold.

6. **Output** - The final calibrated difficulty parameters (b_i) for each item, reflecting their relative difficulty, and ability estimates (θ_j) for respondents, on a shared scale.

In summary, the Rasch model ensures that the resulting scale is invariant across populations and items, meaning item difficulties remain consistent regardless of the sample used (Baker & Kim, 2004). The EM algorithm allows for efficient and robust parameter estimation, even with incomplete response data or large datasets. In the Rasch model, item difficulty can be thought of as the height of a mountain: no matter who is climbing it, the height remains constant. The model separates the inherent difficulty of the item (the mountain's height) from the ability of the climber to ensure that difficulty is measured consistently, regardless of who attempts it. This combination of the Rasch model and the EM algorithm provides a statistically rigorous framework for item calibration, enabling precise measurement of both item difficulty and respondent ability on a common scale. In practical terms, all items that are calibrated together using the Rasch model are placed on a common scale. This enables meaningful comparisons to be made across items and populations. As a result, the item calibration component of the proposed data-driven process of setting benchmarks for RC precursor skills can be summarized by the following:

- **Concurrent Calibration:** All subtasks, including RC, ORF, LC, LS, and others, were calibrated together. This approach ensured that all items were placed on the same scale, allowing for meaningful comparisons between items and subtasks.
- **Rasch Model:** The Rasch model was selected for its ability to create invariant measurement scales. It estimates item difficulty parameters (b_i) and respondent abilities (θ_j) independently of the specific population and items used, ensuring consistency and reliability across different contexts.
- **Context-Specific Calibration:** By stratifying the data and calibrating all subtasks together, the study accounted for contextual variations (e.g., differences in educational systems, languages, and assessment designs) while maintaining the statistical rigor of a unified measurement model.

In this approach, timed tasks such as ORF are treated as subtests consisting of k items, where k represents the passage or task length. Each sequential word in the passage is considered an individual item, and the ability to reach that item as well the correctness of its pronunciation determines the item's score. For example, in a passage of 100 words, the task would be treated as 100 separate items. If a student reads 15 words correctly, then 70 words incorrectly, and does not attempt the final 15 words, the scoring would be as follows:

- 15 items scored as correct
- 70 items scored as incorrect
- 15 items recorded as unattempted (incorrect)

This approach is generally recognized as standard practice, particularly in detailed scoring instructions available for the most common of the oral assessments, and further validated during the TAG meeting in May 2024 ([Early Grade Reading Assessment \(EGRA\) Toolkit | Education Links](#)).

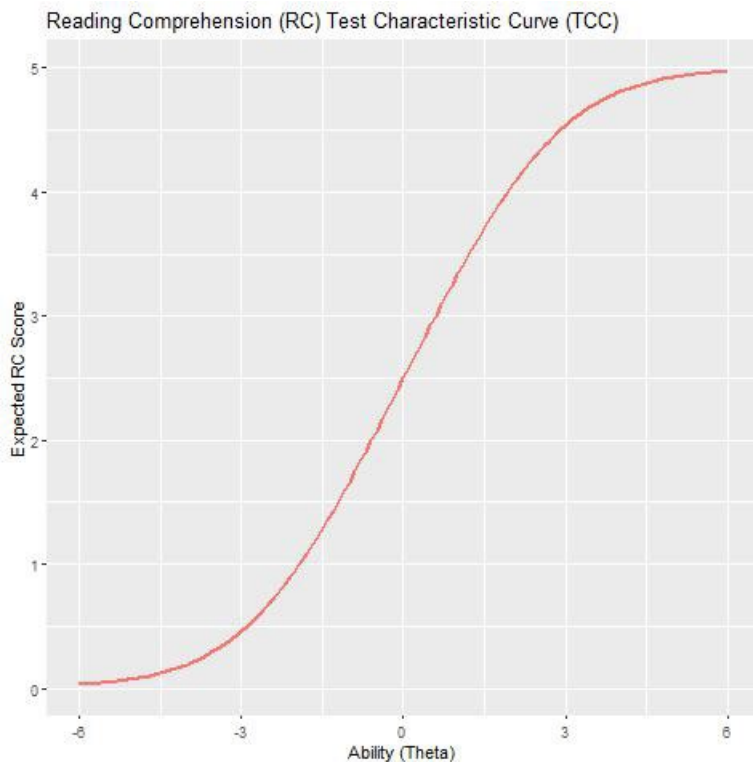
While each word, letter, or syllable is not truly an individual item, and the timed task is conceptually a single item, this representation is necessary for analysis using the Rasch model. The Rasch model requires dichotomous data, so breaking the task into item-level scores for this study allowed timed tasks to be placed on the same scale as other precursor skills that are scored dichotomously. This approach ensured a robust and scalable calibration process applicable across the diverse contexts represented in the datasets. The resulting calibrated item parameters provided a reliable basis for setting benchmarks, analyzing foundational literacy skills, and examining cross-contextual trends in literacy outcomes. For comparison, methods for calculating benchmarks using continuous outcomes based on overall passage performance are also presented below (see Classical Approach for Estimating Accuracy and Benchmarks).

Test Characteristic Curve (TCC)

The next step in the Rasch model approach is to calculate Test Characteristic Curves (TCCs) for each subtask. The TCC represents the relationship between the latent ability (θ) and the expected total test score. It aggregates the probability of correct responses across all items in a test, providing insight into how the test measures ability levels. Below is an example of how a TCC is represented graphically using simulated data for the RC subtask of a typical foundational literacy assessment using 5 comprehension questions.³

³ Note that this 5-question approach was fairly standard practice prior to the requirement, as per the Criteria discussed in this paper, that there be a minimum of 10 comprehension questions. For that reason, this is how the databases used for this paper are structured, and thus this format is used for the paper's examples.

Figure 1. Example of a Reading Comprehension (RC) Test Characteristic Curve (TCC)




To calculate the TCC for a subtask, the difficulty parameters b_i obtained from item calibration are utilized. As stated in the previous section, the Rasch model calculates the probability of a correct response to an item such that:

$$P(X_{ji} = 1) = \frac{e^{1.7(\theta_j - b_i)}}{1 + e^{1.7(\theta_j - b_i)}}$$

where θ_j is the ability level of person j and b_i is the difficulty parameter of item i and the scaling factor 1.7 (Lord, 1980). The TCC is then calculated by summing the expected probabilities of correct responses for all k items within the subtask along a theoretical range of ability levels (θ) such as from -6 to 6:

$$TCC(\theta) = \sum_{i=1}^k P(X_{ji} = 1).$$

In the example above, the simulated difficulty parameters (b) for each item (i) are $b_1 = -2$, $b_2 = -1$, $b_3 = 0$, $b_4 = 1$, and $b_5 = 2$. Each b is the difficulty parameter from the Rasch model equation shown above which calculates the probability of a correct response to an item. In the Rasch IRT model, difficulty parameters indicate how challenging an item is for test-takers. The values typically range from negative to positive, with lower values representing easier items and higher values indicating more difficult items.



For example, an item with a difficulty of -2 would be not very difficult, meaning most students can answer it correctly, while an item with a difficulty of +2 would be much more difficult, answered correctly only by higher-ability students. A difficulty value around 0 suggests a medium-difficulty item that about half of the test-takers are expected to answer correctly. In practice, these values help ensure that test items are appropriately distributed across skill levels, allowing for a more accurate measurement of student ability.

The number of reading comprehension items, however, is not required to be fixed at five items. Increasing the number of reading comprehension items in an EGRA assessment would likely make the TCC steeper and extend its range. With more items, the total test score would span a wider scale, allowing for finer distinctions between different ability levels. Additionally, increasing the number of items would reduce measurement error by providing a more stable estimate of student ability, as each additional item contributes more information to the overall test. However, the impact would depend on the difficulty levels of the new items—if they are evenly distributed across the ability range, the test would better differentiate students across the spectrum, but if clustered at a certain difficulty level, the test’s sensitivity might shift. Overall, a longer test would provide a more detailed picture of student comprehension while potentially altering the way ability estimates align with scoring interpretations.

Conversely, decreasing the number of reading comprehension items would have the opposite effect, potentially flattening the TCC and narrowing its range. With fewer items, the total test score would provide less granularity in distinguishing between ability levels, making it harder to differentiate students, particularly in the middle of the ability distribution. A shorter test would also increase measurement error, as each item would carry more weight, making ability estimates more sensitive to individual responses. If the remaining items do not adequately cover the range of difficulty, the test may become less effective at assessing students at the lower or higher ends of the spectrum. While a reduced item count may improve efficiency and administration time, it could come at the cost of precision and reliability in measuring comprehension ability.

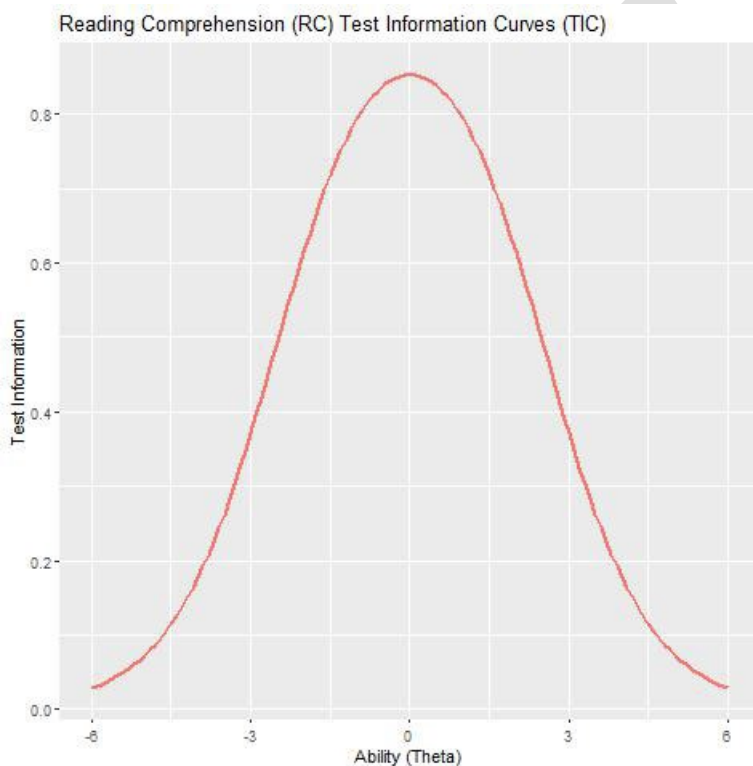
Since foundational literacy assessments are designed for early learners who are still developing literacy skills, it is essential to consider test length when adjusting the number of comprehension items. Younger students typically have limited attention spans and cognitive stamina, making longer tests more likely to cause fatigue, disengagement, or rushed responses, which could introduce additional measurement error. Although increasing the number of items can enhance the reliability of ability estimates, this must be balanced against the practical limitations of test-takers to ensure assessment accurately reflects their true skills. A test that is too long may disadvantage students with lower stamina, potentially underestimating their comprehension ability simply due to test exhaustion. Therefore, decisions about item count should carefully consider both psychometric benefits and the

cognitive demands placed on young learners to maintain validity and fairness in the assessment.

Test Information Curve (TIC)

Once the TCC has been calculated for each subtask, the next step is to calculate the Test Information Curve (TIC) for each subtask. The TIC quantifies the precision of a test at different ability levels (θ). It is derived by summing the item information functions (IIFs) for all items, where item information is a measure of how well an item discriminates between individuals of varying ability levels. Below is an example of how a TIC is represented graphically using simulated data for the RC subtask of a thus-far typical foundational literacy assessment. The simulated data used here is the same as for the TCC.

Figure 2. Example of a Reading Comprehension (RC) Test Information Curve (TIC)



In the Rasch model, the IIF for an individual item is defined as:

$$I_i(\theta) = P(X_{ji} = 1) \cdot (1 - P(X_{ji} = 1)),$$

where $P(X_{ji} = 1)$ is the Rasch probability of a correct response to item i at ability level θ (Birnbaum, 1968). Note that the IIF will be maximized when $P = 0.5$. This occurs because the IIF reaches its maximum value when the probabilities of a correct and incorrect response are equal, i.e., when the outcome is most uncertain. Interpretively, this means that an item provides the most information about a respondent's ability when (without taking a respondent's ability into account) there is a 50% chance of answering it correctly or

incorrectly. An equal probability of answering an item correctly or incorrectly means that the item is as sensitive as possible to differences in ability levels, making it optimal for distinguishing between respondents that are near the item's difficulty level. The TIC for the entire test (in this case, subtask) is then expressed as:

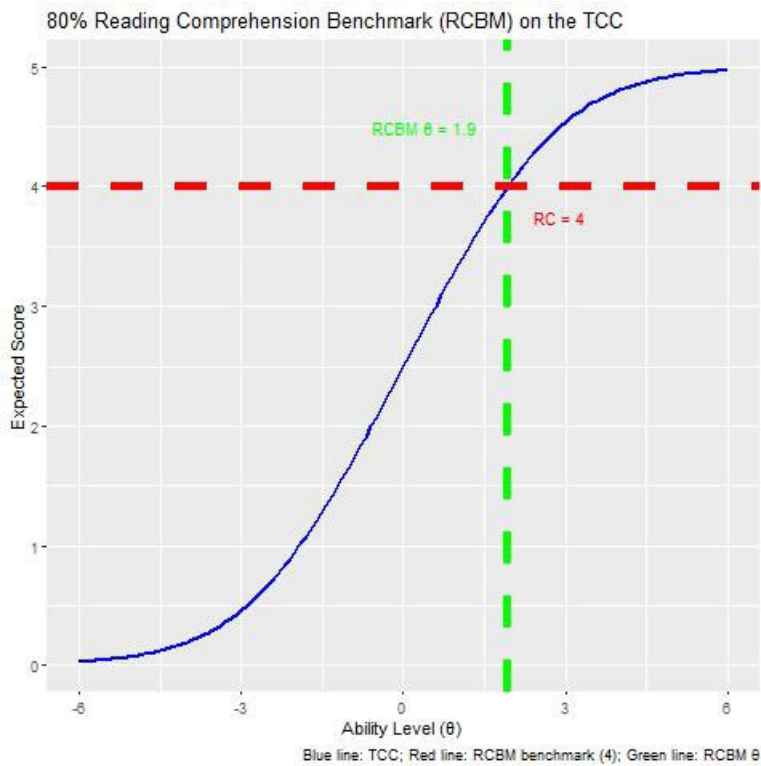
$$TIC(\theta) = \sum_{i=1}^k I_i(\theta)$$

where k is the total number of items in the test/subtask (Samejima, 1977). In the above example, the largest amount of information is found where ability level (θ) is approximately equal to the item difficulty (b), which is 0 in this case. This indicates the highest measurement precision at that point on the ability scale for this subtask. This occurs because the items in this example are symmetrically distributed around $b = 0$. As a result, the test is most informative around $\theta = 0$, where the respondent's ability matches the average item difficulty, i.e. where there is the most sensitivity to differences in ability levels and the highest measurement precision.

Estimating Reading Comprehension Benchmark (RCBM) Ability

Once the item parameters have been estimated and the TCCs for each subtask have been calculated, the next step in the initially proposed data-driven approach to benchmarking RC precursor skills is to identify the ability levels (θ) on the RC TCC that correspond to RC scores of 60% and 80% correct. The Reading Comprehension benchmark (RCBM) ability levels corresponding to 60% and 80% accuracy are determined by identifying the ability level (θ) at which the expected score on the TCC is closest to the benchmark score for 60% and 80% correct (on a test with five RC items, this corresponds to three items correct for 60% RCBM and four items correct for 80% RCBM). Below is a graphical representation of this process for an 80% correct RCBM on an assessment that has five RC items, using the same simulated data as in the prior examples.

Figure 3. Example of Estimating Reading Ability at 80% RCBM via the TCC



In the example above, the RCBM is set at 80% accuracy for a RC subtask with five items, which means that the benchmark score would be four correct items out of five total items. As a result, the RCBM for 80% correct is the θ -value where the difference between the expected score (from the TCC) and the benchmark score (4) is minimal. Graphically, the process starts with the given expected score the vertical axis. One then reads to the right from there using the red dotted one until that line meets the TCC, and then one reads from that intersection down to the horizontal axis, to thus reading how the number of correct responses on the vertical axis translates into the θ -value on the horizontal axis. In the example illustrated above, the corresponding θ -value for 80% RC accuracy is approximately 1.9. That is, one is reading from a criterion-referenced benchmark for questions that are answered correctly to an estimated measure of latent ability (θ).

Estimating Reading Comprehension Standard Error (RCSE)

Next, the standard error (SE) of the RCBM ability level (θ) is obtained by first locating the test information value ($I(\theta)$) which corresponds to θ . Traditionally, SE is the standard deviation of the sample-based estimate of a population parameter, but in IRT, it represents the uncertainty in estimating a student's ability level (θ) based on their test performance. Instead of reflecting variability across repeated samples, SE in this context indicates how precisely the test measures ability at different points along the θ scale. A lower SE means a more precise estimate, while a higher SE suggests greater uncertainty in assessing a student's true ability. SE is directly tied to test information ($I(\theta)$), which quantifies how well the test differentiates

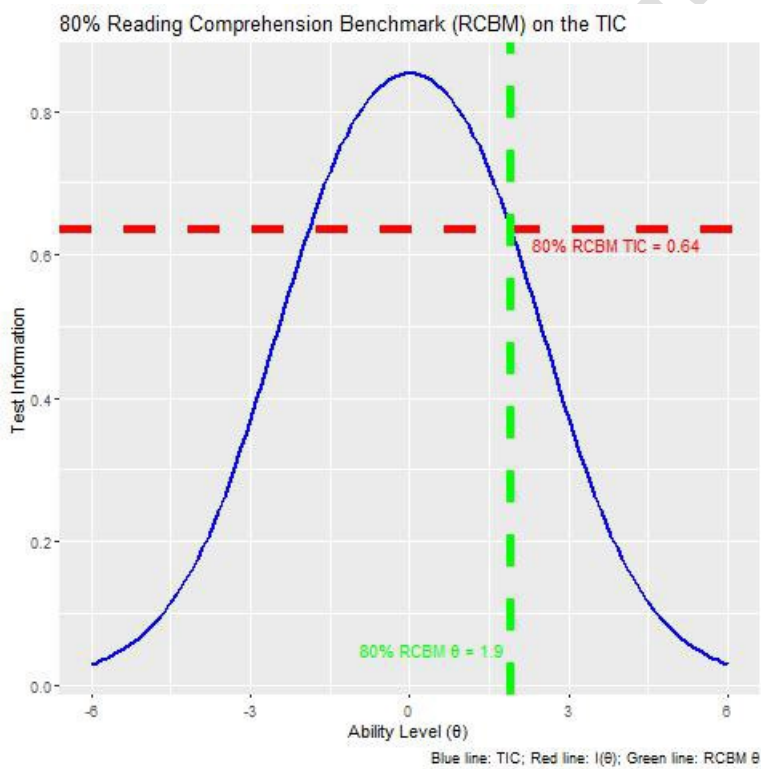
ability levels—higher test information leads to lower SE, meaning greater confidence in the ability estimate.

Once the test information value for θ has been obtained, the SE for θ is calculated by the inverse of the square root of $I(\theta)$ as shown in the following equation:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

To illustrate the process of locating $I(\theta)$, below is a graphical representation using the same simulated data from the above example for estimating θ using the TCC.

Figure 4. Example of Estimating Test Information via the TIC for 80% RCBM

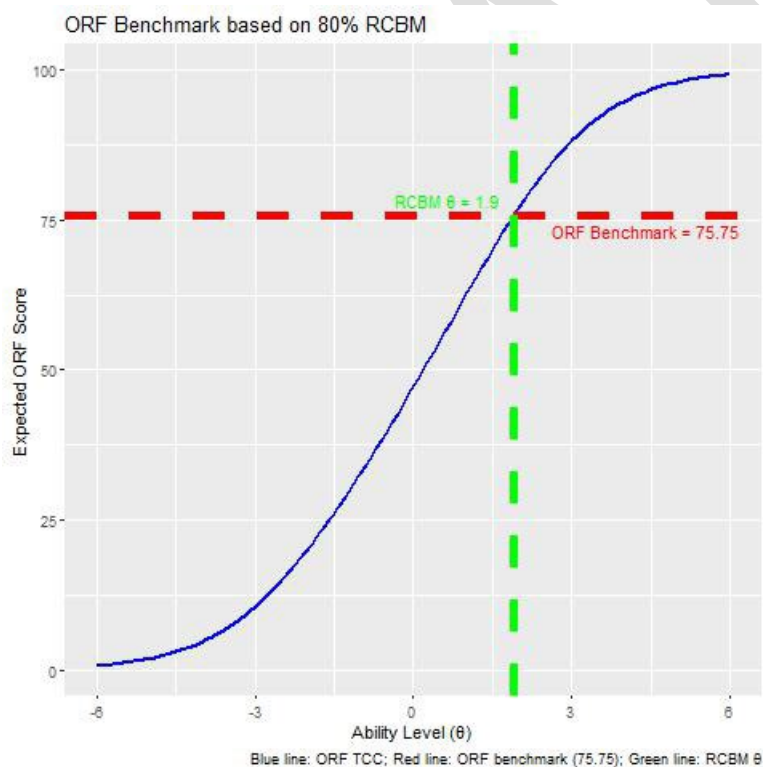


In the example above θ is 1.9, which, according to the TIC, means that the test information value for that θ is 0.64 ($I(\theta_{RC}) = 0.64$). This means that the corresponding SE for a θ value of 1.9 is 1.25, because $1/\sqrt{0.64} = 1.25$. Therefore 1.25 is the SE for the 80% RCBM ($RCSE = \frac{1}{\sqrt{0.64}} = 1.25$). In other words, the location on the θ ability scale of the 80% RCBM (1.9) corresponds to a SE of 1.25 according to the TIF for Reading Comprehension. To trace these steps graphically using the figure above, one begins at 0 on the horizontal x-axis, finds location of the ability level (θ) of 1.9, then reads up along the dotted green line until it intersects the TIC, and reads to the left along the dotted red line. This is then used to estimate the SE, as per the equation.

Estimating Benchmarks for Precursor Skills

Once the corresponding locations on the ability scale (θ) have been determined for the RC benchmarks of 60% and 80% correct, these ability levels are used to identify the expected scores on the TCCs of the precursor skills. These expected scores represent the calibrated benchmarks for the precursor skills, aligning them with the RC proficiency levels. By translating RC ability levels into equivalent performance expectations for each subtask, this approach ensures consistency across different measures of foundational literacy and, as a result, provides a unified framework for interpreting student performance across subtasks, using the same measure of latent ability θ for all sub-tasks such as RC, ORF, etc., while maintaining a direct connection between all the precursor skills jointly and the RC scale. Below is a graphical representation of this process for using the 80% RCBM in the previous examples to estimate the corresponding benchmark for ORF using simulated data. It is also important to note that earlier attempts to set benchmarks using classical regression techniques do not base themselves on a unified and coherent view of the test-takers' latent or inherent ability and assume all items on the RC scale and on the scale of the precursor skills are of equal difficulty.

Figure 5. Example of Estimating ORF Benchmarks for 80% RCBM



In the example above, the 80% RCBM estimate for θ (1.9) corresponds to an expected score of 75.75 on the TCC for ORF. As a result, the benchmark for ORF according to a RCBM of 80% is approximately 76. In this case, one can find the location of the 80% RCBM θ by the location of the vertical green line on the x-axis and trace it to the blue TCC line. The point at which they intersect is then where one can trace the horizontal red line to the y-axis to find the aligned ORF benchmark.

Estimating the Subtask Confidence Interval

A confidence interval (CI) was constructed around the benchmark estimates for each subtask. Initially, these intervals were calculated by adding and subtracting the quantity $1.96 \times RCSE$ from RCBM θ and finding the corresponding SE for those lower and upper θ values on that subtask's TIC. Multiplying the SE by 1.96 is derived from the properties of the standard normal distribution, for which about 95% of the data falls within 1.96 standard deviations of the mean (thus corresponding to a 95% CI). As such, the upper and lower bounds for the confidence interval for the f estimate of the ability parameter can be calculated by:

$$\theta_{RCBM} \pm 1.96 \times RCSE,$$

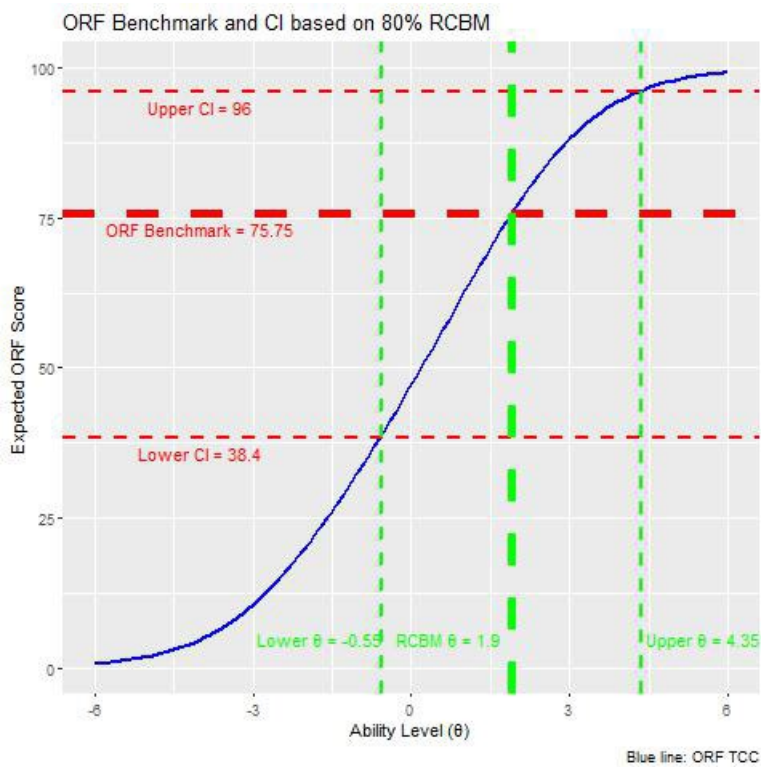
Where $\theta_{lower} = \theta_{RCBM} - 1.96 \times RCSE$ and $\theta_{upper} = \theta_{RCBM} + 1.96 \times RCSE$ and the estimates for θ_{lower} and θ_{upper} cannot exceed the minimum and maximum range of θ . If the estimates for the bounds exceed the minimum or maximum value of θ being estimated, the respective bound is equal to that minimum/maximum which is being exceeded.

The process for estimating the subtask confidence interval is illustrated below, using the simulated data values from the previous examples for RC and ORF. According to the above equation the lower and upper RCBM θ bounds are equal to:

$$1.9 \pm 1.96 \times 1.25.$$

As a result, the lower θ bound of RCBM equals -0.55 while the upper θ bound of RCBM equals 4.35. Next, to estimate the confidence interval for ORF, those bounds are located on the ORF TCC and subsequently their corresponding expected scores. In this example, the corresponding values are 38.4 for the lower bound and 96.00 for the upper bound. Below is a graphical representation of the process.

Figure 6. Example of Estimating ORF Benchmark Confidence Interval for 80% RCBM




Note that because of the method used, where the CI for the ORF is not developed on the ORF itself but as a projection of the CI for the ability level, the CI for the ORF is not symmetrical.

Addressing Special Cases in Benchmark Estimation: Practical Considerations

In foundational literacy assessments, benchmarks and confidence interval bounds for subtasks occasionally approach the limits of the ability scale (θ), either nearing 0 or exceeding the upper range of 6.0. This can happen due to factors like small sample sizes, high performance variability, or disproportionate scoring patterns. For example, benchmarks tied to higher RC proficiency levels (e.g., 80%) may exceed $\theta = 6.0$ if most students perform well below the expected range, leaving insufficient data to support higher estimates. Similarly, benchmarks approaching 0 reflect scores asymptotically nearing, but not reaching, zero due to the infinite nature of the ability scale. Expanding the scale beyond -6.0 to 6.0 is not recommended, as measurement reliability declines significantly outside this range, and such extreme scores are rare in real-world data. By constraining the ability scale to this range, the study ensures reliable, interpretable, and practical results that remain grounded in the realities of literacy assessment.

When using real data, it is important to recognize that in some cases the location of θ for either a subtask benchmark or a confidence interval bound may correspond to an expected score of 0. This can occur due to factors such as small sample size, high variance in student performance on the relevant subtask or RC, an overrepresentation of low scores relative to the possible total score on the subtask or RC, or a combination of these factors. In these



scenarios, while the estimated benchmark or confidence interval bound may appear to be 0, it is not exactly 0. Instead, the expected score is asymptotically approaching 0 because the ability scale is theoretically infinite, extending towards negative infinity. This effectively brings the expected score closer to 0 without ever reaching it. In practical terms, this outcome indicates that the benchmark should be interpreted as 1, as one item is the minimum number of items that can be scored correctly—the lowest achievable score across all subtasks.

Similarly, estimated benchmarks or confidence interval bound for a subtask may also occasionally exceed the upper limit of the theta scale as well. Such cases may occur when an estimated θ value for a benchmark or confidence interval exceeds the estimated upper limit of theta (which is 6.0 in the present study). This issue can arise due to an interaction between pre-determined RCBM levels and factors such as small sample sizes, low sample performance on RC or the relevant subtask, or a combination of these factors. For example, in this study RCBMs were set at 60% and 80%, which typically corresponds to three or four correct items out of five RC items on an assessment, respectively. If most students score only one or two items correctly on RC, the location of θ for the 80% RCBM (four correct items) would most likely exceed 6.0, because the TCC would only be able to show data for answering at most two items correctly.

At first glance, it may appear that expanding the estimated range of the ability scale would be a viable solution for these situations. However, this approach is not recommended due to practical limitations in measurement accuracy. In most cases, person-level ability estimates for any given subtask will not exceed ± 6.0 , for either high- or low-performing individuals or high- or low-performing sample populations. Beyond ± 6.0 estimated ability SE increases significantly, which greatly reduces the reliability of measurements beyond those points. Foundational literacy assessments are not designed to measure extreme outliers, although such results may be objectively valued when they occur.

These assessments aim to evaluate literacy proficiency in populations that closely reflect real-world distributions. Extremely high scores, far beyond the expected range, are statistically improbable in real data and would most likely skew the results if they were to occur and be included in analysis of the sample. Such outliers, while noteworthy, would provide limited utility for generating broadly applicable insights. As a result, the study recommends constraining the estimated ability scale to a range of -6.0 to 6.0. This will ensure that the results remain reliable, interpretable, useful for practical applications, and grounded in empirical data.

Classical Approach for Estimating Accuracy and Benchmarks

In this study, the classical approach for setting benchmarks was incorporated as a comparative framework alongside the proposed data-driven IRT method, ensuring a comprehensive evaluation of literacy assessment strategies.

Timed Task Accuracy

For timed tasks, accuracy measures a reader’s ability to decode and pronounce words and letters correctly, highlighting their mastery of foundational reading skills such as phonics. Accuracy focuses exclusively on the correctness of reading, independent of the amount of time used by the student to read the passage. As a result, the classical approach for measuring timed task accuracy provides a straightforward measure of a student’s ability to recognize and pronounce words accurately. Using ORF as an example, the calculation of ORF accuracy (*Accuracy*) can be represented with the following equation:

$$Accuracy = \frac{\text{Correct Words Read}}{\text{Total Words in Passage}} \times 100$$

For example, if a student read a passage which contained 50 words and correctly read 45, their accuracy score would be calculated as follows:

$$Accuracy = \frac{45}{50} \times 100 = 90\%$$

High accuracy is a prerequisite for comprehension, as readers must decode text reliably before focusing on its meaning.⁴ However, accuracy alone does not capture the full complexity of proficient reading.

In addition to accuracy, foundational literacy assessments also emphasize timed tasks. These tasks do not always impose a strict time limit but instead measure the time required for a student to read a passage, word list, or letter-reading task. Such timed precursor skills, including ORF, are crucial indicators of reading proficiency, capturing a student’s ability to read text accurately, efficiently, and with proper expression.

While accuracy and fluency are interrelated, they measure distinct components of reading performance—both of which are essential for achieving RC and fostering literacy development (Fuchs et al., 2001). Therefore, when establishing benchmarks for RC precursor skills, fluency must be considered a complementary dimension of literacy. (Note some oral assessments may not explicitly measure fluency and may instead focus solely on accuracy. In such cases, the approach defaults to accuracy).

Classical Approach for Foundational Literacy Subtask Benchmark Estimation

The classical approach for benchmark estimation on foundational literacy assessments uses the following two-step procedure which can be applied to RC and RC precursor skills:

⁴ The standard threshold is 95%.

1. Obtain the mean ($\bar{X}_{subtask}$) total score of all RC precursor skills and the standard error (SE) of all subtasks from students that achieved the set RCBM score.
 - For example, if the RCBM was 80% on a five-item test, sample would be filtered to just those students that answered exactly four items correctly on RC and the mean score and SE for each subtask would be obtained.

2. The confidence interval (CI) for each subtask is calculated by adding and subtracting the quantity $1.96 \times SE$ from the mean score of each subtask. Multiplying the SE by 1.96 is derived from the properties of the standard normal distribution, for which about 95% of the data falls within 1.96 standard deviations of the mean (thus corresponding to a 95% CI). In IRT, this is commonly used when reporting confidence intervals for estimated parameters like ability (θ).
 - The formula for the SE of each subtask is $SE = \frac{SD}{\sqrt{n-1}}$, where SD is the standard deviation of the subtask total score (for students at the RC benchmark) and n is the sample size.

As a result, the mean subtask total score ($\bar{X}_{subtask}$) of students that met the RCBM is the new subtask benchmark ($BM_{subtask}$) for those precursor skills, and the classical approach formula (after filtering the data for students that did not meet the RCBM) for setting benchmarks for RC precursor skills is:

$$BM_{subtask} = \bar{X}_{subtask},$$

And the formula for calculating CIs for all subtasks is:

$$BM_{subtask} \pm (1.96 \cdot SE_{subtask}).$$

Assessment Data for Initial Benchmark Analysis

Four EGRA data sets were selected for the initial benchmark analysis. These included:

- EGRA Country 4 Arabic Grade 2
- EGRA Country 6 Arabic Grade 2
- EGRA Country 13 Chitonga Grade 2
- EGRA Ghana English Grade 2

Confirmatory Analysis – Correlations and Principal Components Analysis (PCA)

Analysis was conducted to confirm the unidimensionality of the selected assessments. The results of Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) confirmed the datasets' suitability for Principal Components Analysis. The high chi-square values and statistically significant p-values demonstrate strong correlations among the variables, while the KMO values indicate excellent sampling adequacy. This lays a solid foundation for identifying latent components and reducing dimensionality in the dataset. Further exploration of eigenvalues, explained variance, and loadings is recommended to interpret the underlying structure effectively.

Table 1. Bartlett's Test of Sphericity

Assessment	Bartlett.chisq	Bartlett.p.value	Bartlett.df
EGRA Country 4 Arabic Grade 2	17861	<0.001	21
EGRA Country 6 Arabic Grade 2	9182	<0.001	15
EGRA Country 13 Chitonga Grade 2	5072	<0.001	15
EGRA Ghana English Grade 2	4572	<0.001	10

Bartlett's test of sphericity was conducted to determine whether the variables were sufficiently correlated to justify the application of Principal Components Analysis (PCA). Chi-square values ranged from 4,572 to 17,861, depending on the variable or region. P-values were consistently less than 0.001 (or even 0.0001), indicating statistical significance and confirming the appropriateness of PCA for the dataset. These results suggest that the variable exhibits the necessary intercorrelations required for effective dimensionality reduction and analysis.

Table 2. Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Subtask	EGRA Country 4 Arabic Grade 2	EGRA Country 6 Arabic Grade 2	EGRA Country 13 Chitonga Grade 2	EGRA Ghana English Grade 2
Overall	0.81	0.81	0.84	0.80
Listening Comprehension	0.83	0.85	0.93	0.89
Letter Sound	0.78	0.83	0.93	0.84
Syllable Sound	0.88		0.83	
Invented Word	0.82	0.83	0.82	0.83


Subtask	EGRA Country 4 Arabic Grade 2	EGRA Country 6 Arabic Grade 2	EGRA Country 13 Chitonga Grade 2	EGRA Ghana English Grade 2
Oral Reading Fluency	0.75	0.77	0.80	0.74
Reading Comprehension	0.73	0.80	0.84	0.76
Silent Reading Comprehension	0.90			

The KMO measure was calculated to assess sampling adequacy both overall and for individual variables. Overall KMO values equaled or exceeded 0.8 for all assessments, indicating excellent adequacy (Kaiser & Rice, 1974). Task-specific KMO values ranged from 0.78 to 0.93 across subtasks and assessments, suggesting robust sampling adequacy for most variables analyzed. These results further validate the suitability of the data for PCA. It is important to note that the KMO test is a measure of sampling adequacy in the sense of the correlations among items, not the number of students (sample size). It assesses whether the patterns of correlations among variables (items) are suitable for factor analysis by determining the proportion of variance that might be common variance (i.e., shared among items) rather than unique variance. A high KMO value (greater than or equal to 0.80) suggests that the items have enough shared variance for factor analysis, while a low KMO value (less than 0.50) indicates that the items are poorly correlated, and factor analysis may not be appropriate.

Table 3. Principal Components Analysis (PCA)

Variable	Subtask	EGRA Country 4 Arabic Grade 2	EGRA Country 6 Arabic Grade 2	EGRA Country 13 Chitonga Grade 2	EGRA Ghana English Grade 2
Eigenvalue	Overall	4.41	3.38	3.96	3.22
Variance Explained by Factor 1	Overall	0.63	0.56	0.66	0.64
Factor Loading on Factor 1	Listening Comprehension	0.48	0.45	0.32	0.63
	Letter Sound	0.64		0.73	0.76
	Syllable Sound	0.88		0.92	
	Invented Word	0.87	0.85	0.94	
	Oral Reading Fluency	0.92	0.91	0.93	
	Reading Comprehension	0.85	0.83	0.85	
	Silent Reading Comprehension	0.81			

The results of the PCA shown in the table above indicate that the eigenvalues exceed 1 for all assessments, confirming the first principal component captures a substantial portion of the



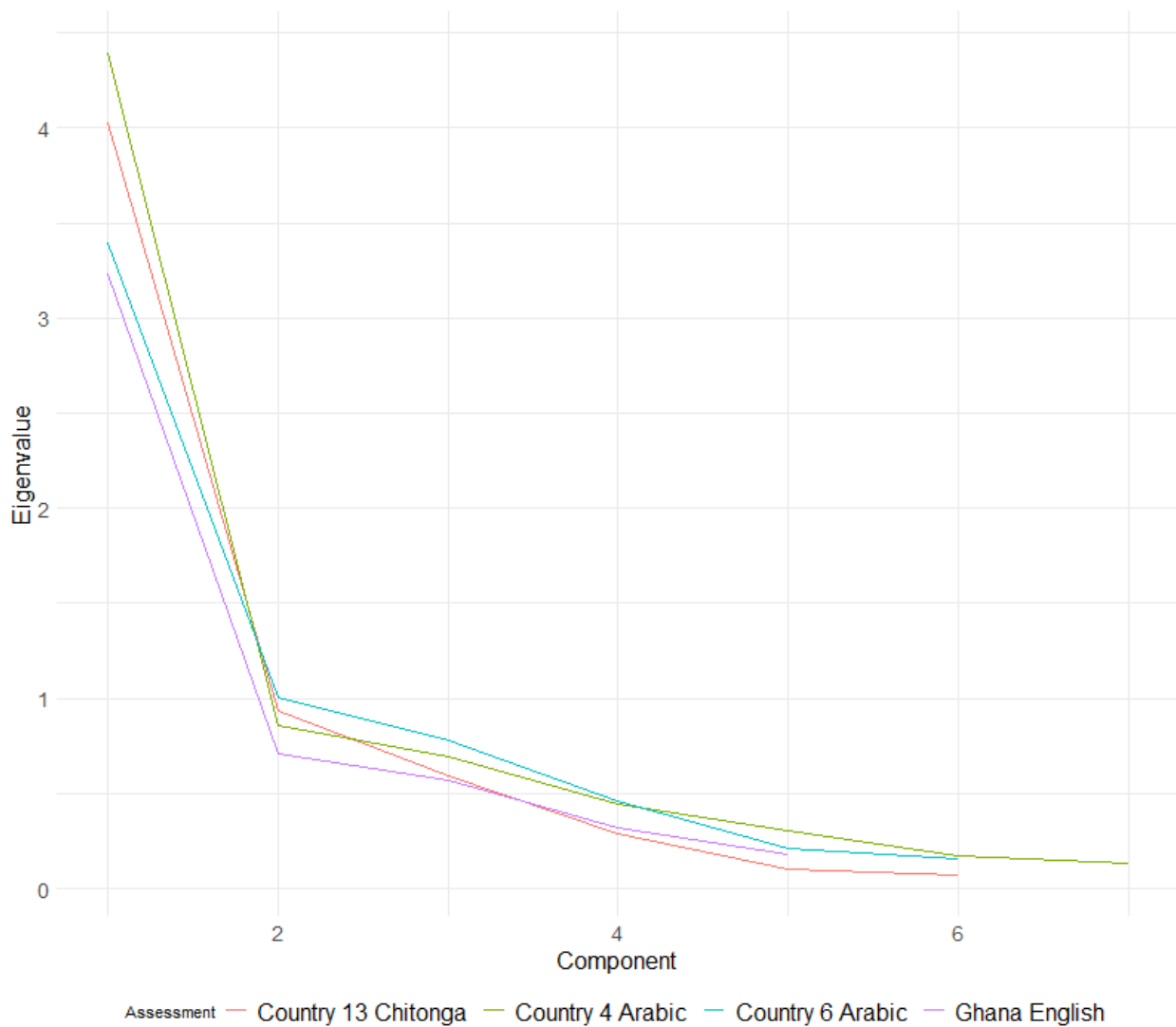
variance, consistent with the patterns observed in the scree plots below. The proportion of variance explained by the first factor (Component 1) varies slightly across assessments:

- **EGRA Country 4 Arabic Grade 2: 63%**
- **EGRA Country 6 Arabic Grade 2: 56%**
- **EGRA Country 13 Chitonga Grade 2: 66%**
- **EGRA Ghana English Grade 2: 64%**

These values indicate that Component 1 explains more than half of the total variance in all datasets, suggesting a dominant underlying pattern in student performance across the assessments (Everitt & Dunn, 2001). ORF and Invented Word consistently exhibited the highest loadings, suggesting they are key contributors to the primary dimension of variance across assessments. The results demonstrate strong unidimensionality across all datasets, with the first component consistently explaining the majority of variance and tasks loading strongly onto this component. This unidimensionality makes the data well-suited for analysis using the Rasch IRT model, which assumes a single latent trait underlies observed performance. The Rasch model would provide robust, interpretable measures of student ability and item difficulty across these diverse assessments.

Below the scree plot generated from the PCA for the four assessments is displayed. The plot visualizes the eigenvalues associated with each principal component for these assessments. A pronounced decline in eigenvalues is evident from Component 1 to Component 2 across all assessments. This indicates that Component 1 captures most of the variance in the dataset. The scree plot shows a noticeable “elbow” around Component 2 for most assessments, suggesting that the first two components are likely sufficient to explain most of the variance and components beyond this point contribute relatively little to the total variance. The trends in eigenvalues are also consistent across all four assessments, indicating a similar structure in the data across these regions and tasks.

Figure 7. Scree Plot from Principal Components Analysis (PCA)



As a result, the results of the PCA and scree plot reveal that the data is largely unidimensional and that a dominant underlying component explains the majority of variance in student performance across the four assessments. The unidimensional structure revealed by PCA highlights the dominance of a single latent construct across student tasks and regions, a latent trait can be interpreted as Foundational Reading Skill. This trait aligns with key literacy skills, as evidenced by the consistently high loading values for tasks like ORF. The suitability of the Rasch IRT model ensures that this data can be effectively calibrated and analyzed, supporting scalable and valid assessment frameworks.

Table 4. Item/Test Correlation by Assessment

Subtask/Item	EGRA Country 4 Arabic	EGRA Country 6 Arabic	EGRA Country 13 Chitonga	EGRA Ghana English
Invented Word (timed task)	0.83	0.86	0.93	0.83
Letter Sound (timed task)	0.79	0.95	0.82	0.83

Listening Comprehension Item 1	0.25	0.21	0.16	0.43
Listening Comprehension Item 2	0.18	0.30	0.15	0.36
Listening Comprehension Item 3	0.28	0.23	0.16	0.42
Listening Comprehension Item 4	0.18	0.26	0.26	
Listening Comprehension Item 5	0.29		0.14	
Oral Reading Fluency (timed task)	0.84	0.91	0.89	0.92
Reading Comprehension Item 1	0.41	0.47	0.48	0.55
Reading Comprehension Item 2	0.32	0.42	0.61	0.60
Reading Comprehension Item 3	0.12	0.19	0.46	0.43
Reading Comprehension Item 4	0.20	0.33	-0.03	0.39
Reading Comprehension Item 5	0.21	-0.03	-0.27	0.25
Reading Comprehension Item 6		0.27		
Reading Comprehension Item 7		0.20		
Syllable segmentation Item 1		0.25		
Syllable segmentation Item 2		0.25		
Syllable segmentation Item 3		0.25		
Syllable segmentation Item 4		0.30		
Syllable segmentation Item 5		0.25		
Syllable segmentation Item 6		0.23		
Syllable segmentation Item 7		0.15		
Syllable segmentation Item 8		0.22		
Syllable segmentation Item 9		0.11		
Syllable segmentation Item 10		0.17		
Syllable Sound (timed task)	0.90		0.96	

The table provides item/test correlations for various subtasks across four regions: EGRA Country 13 Chitonga, EGRA Country 4 Arabic, EGRA Country 6 Arabic, and EGRA Ghana English. Several items demonstrate strong correlations (≥ 0.8) across multiple regions, indicating their consistent alignment with the test construct. For example, Invented Word Item 1 shows correlations ranging from 0.80 to 0.93, suggesting it is a reliable indicator of the latent ability across all regions. Similarly, LS Item 1 exhibits high correlations (0.79–0.95), confirming its robustness, while ORF Item 1 has consistently strong correlations (0.84–0.92), highlighting its effectiveness in capturing the intended construct.

Moderate-performing items, with correlations between 0.4 and 0.7, include RC Item 1 and Item 2, which show meaningful contributions to the tests, with correlations ranging from 0.32 to 0.61. Additionally, SS Items 1–6 demonstrate modest correlations (0.17–0.30), suggesting their performance, while meaningful, is less substantial compared to high-performing items. In contrast, some items performed poorly, with correlations below 0.3 or showing inconsistencies across regions. For example, LC Items generally show weak correlations (0.16–0.43), with the highest observed in Ghana English Grade 2. These results suggest these items may not align well with the primary test construct measured by the assessment. Similarly, RC Items 4 and 5 exhibit negative or near-zero correlations in some regions, such as -0.27 for RC

Item 5 in Chitonga, indicating potential issues. Overall, the analysis highlights several robust items, such as ORF, INW, and LS, which perform consistently across regions. However, weaknesses in certain items that suggest targeted revisions are necessary to improve alignment with the underlying test construct and ensure both validity and fairness in the assessments.

Having established the suitability of the four assessments against, this study then conducted the initial benchmark analysis on all four data sets.

Results of the Initial Rasch Accuracy IRT Model and Classical Approach

For all IRT analyses, item calibration and estimation of Rasch model parameters was conducted via the free/open-source software R (R Core Team, 2024), specifically the R package ‘mirt’ (Chalmers, 2012). Below are the results of the initial Rasch Accuracy IRT Model and classical approach analysis:

Table 5. 60% RC-based Benchmarks set by the Classical Approach

Assessment	Subtask	Items	SE	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.61	26	27	28
	Reading Comprehension	5	NA	3	3	3
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	2.58	52	57	62
	Reading Comprehension	7	NA	4	4	4
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	0.78	24	26	27
	Reading Comprehension	5	NA	3	3	3
EGRA Ghana English Grade 2	Oral Reading Fluency	60	3.74	45	52	59
	Reading Comprehension	5	NA	3	3	3

In the table above, the Benchmark column represents the benchmark and is calculated as the average performance for each subtask among students with RC scores equal to the target or threshold. For example, the ORF benchmark for EGRA Country 4 Arabic Grade 2 is 27, meaning that students who scored a 3 on RC had an average ORF score of 27. Similarly, the benchmark for ORF in EGRA Ghana English Grade 2 is 52, indicating the average performance of students with a RC score of 3. This column serves as the reference point for interpreting student performance in relation to the defined RC threshold.

Table 6. 80% RC-based Benchmarks set by the Classical Approach

Assessment	Subtask	Item	SE	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.71	33	34	36
	Reading Comprehension	5	NA	4	4	4
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	5.07	67	77	87
	Reading Comprehension	7	NA	6	6	6
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	1.1	31	33	35
	Reading Comprehension	5	NA	4	4	4
	Oral Reading Fluency	60	4.75	68	77	87

EGRA Ghana English Grade 2	Reading Comprehension	5	NA	4	4	4
----------------------------	-----------------------	---	----	---	---	---

In the table above, the Benchmark column represents the benchmark and reflects the average performance for each subtask among students with RC scores meeting the 80% threshold. For example, the ORF benchmark for EGRA Country 4 Arabic Grade 2 is 34, indicating that students who scored a 4 on RC had an average ORF score of 34. Similarly, for EGRA Ghana English Grade 2, the benchmark for ORF is 77, showing the average performance of students with a RC score of 4. This column provides a key reference for evaluating student performance relative to the 80% RC threshold.

Table 7. 60% RC-based Benchmarks and CIs set by the Rasch Accuracy IRT Model

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	1.03	4	20	37
	Reading Comprehension	5	1.03	0	3	5
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	1.04	19	36	46
	Reading Comprehension	7	1.04	2	4	6
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	1.22	2	16	29
	Reading Comprehension	5	1.22	1	3	4
EGRA Ghana English Grade 2	Oral Reading Fluency	60	1.13	20	49	54
	Reading Comprehension	5	1.13	0	3	4

In the table above, the Benchmark column represents the estimated subtask scores derived from the IRT method, which uses RC theta standard errors (θ_{SE}) to calculate confidence intervals (LCI and UCI) for each subtask. These benchmarks reflect the expected performance of students whose RC ability corresponds to 60% accuracy, considering measurement uncertainty. For example, in EGRA Country 4 Arabic Grade 2, the ORF benchmark is 20, meaning students scoring at the 60% RC threshold are expected to achieve an average ORF score of 20. Similarly, for EGRA Ghana English Grade 2, the benchmark for ORF is 49, with confidence intervals ranging from 20 to 54, illustrating the range of expected scores based on theta uncertainty. This approach incorporates measurement error (θ_{SE}), providing a statistically sound framework for defining benchmarks that align with specific RC proficiency levels. The benchmarks and their confidence intervals allow for nuanced interpretation of subtask performance in relation to RC proficiency.

Table 8. 80% RC-based Benchmarks and CIs set by the Rasch Accuracy IRT Model

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	1.21	6	27	41
	Reading Comprehension	5	1.21	1	4	5
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	1.1	27	46	46
	Reading Comprehension	7	1.1	3	6	6
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	2.13	1	26	29
	Reading Comprehension	5	2.13	1	4	4
	Oral Reading Fluency	60	1.3	23	54	54

EGRA Ghana English Grade 2	Reading Comprehension	5	1.3	1	4	4
-------------------------------	-----------------------	---	-----	---	---	---

In the table above, the Benchmark column represents the expected subtask scores derived using the IRT method, which incorporates the RC theta standard errors (θ_{SE}) to calculate confidence intervals (LCI and UCI). These benchmarks are tied to the 80% RC accuracy threshold, reflecting the average performance of students at this proficiency level while accounting for uncertainty in their theta estimates. For example, in EGRA Country 4 Arabic Grade 2, the ORF benchmark is 27, with a range of scores between 6 (LCI) and 41 (UCI). This suggests that students achieving 80% RC accuracy tend to score, on average, 27 on ORF. In EGRA Ghana English Grade 2, the benchmark for ORF is 54, with confidence intervals spanning 23 to 54, indicating a narrower range of expected performance for students at this proficiency level.

Note that the reason why the SEs are presented as the same for the ORF and RC subtasks is that they are the SEs for the RC benchmarks of 60% and 80% comprehension, and this θ is the same for all sub-tasks, given the effort to estimate all the benchmarks jointly and based on a θ that is the generalized child ability in reading, which, as explained above, is an attractive feature of the IRT method as opposed to the classical method. As explained graphically above, the SEs for the θ are used to create a confidence interval for θ , based on the RC benchmark and RCSE and this interval is projected onto the ORF TCC.

Limitations

The primary limitation of these results from the initial Rasch Accuracy IRT Model approach identified by the TAG in May 2024 was the size of the confidence intervals observed for non-RC subtasks. These broad intervals reduced the precision of the estimated benchmarks, making it challenging to draw accurate conclusions. It was not immediately clear why the CIs using the IRT method should be so much broader than using a classical method. As a result, the TAG recommended developing solutions to reduce the size of these confidence intervals by exploring the suitability of an IRT method based on a binomial approach where each word is considered a separate item, in a task ultimately based on a sense of fluency. Narrower intervals would allow for greater precision and reliability in interpreting benchmarks. This improvement was particularly critical for ensuring robust assessments in non-RC subtasks as the ones being set by the data-driven methods of the approach.

The next section presents the result of the additional analysis conducted as a result of the TAG recommendations.

Section 4: Updated Results from TAG Recommendations

Following the recommendations of the TAG to explore why the novel application of IRT to the analysis of foundational skills in the manner described above resulted in SEs that are so much larger than those procured by the classical method, it was decided to re-think how the SEs of the precursor skills could be estimated in a more realistic manner. This was approached by two methods:

1. A revised approach to the Rasch Accuracy IRT Model, in which the SE for each subtask is based directly on the TIC of that subtask, instead of being based on the TIC of reading comprehension and projected onto the subtask using the SEs of the θ value for RC.
2. A generalized linear mixed-model (GLMM) method for estimating benchmarks.

Rasch Accuracy IRT Model: Revised Approach

The revised approach to the Rasch Accuracy IRT Model is modified at the step for estimating each subtasks SE, whereby SE is calculated by adjusting the RC benchmark ability level (θ) by $\pm 1.96 \times SubtaskSE$, where *SubtaskSE* represents the standard error of the ability estimate for each precursor skill. In other words, in the initial approach only one SE estimate was used for all confidence intervals (based on the RC TIC), while in the revised approach *s* estimates for SE were used for each subtask, where *s* represents the number of precursor skills in the assessment.

Revised Rasch Accuracy IRT Model Results

Table 9. Rasch Accuracy IRT Benchmarks According to 60% RC and Subtask SE


Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.42	12	20	28
	Reading Comprehension	5	1.03	0	3	5
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	0.45	29	36	44
	Reading Comprehension	7	1.04	2	4	6
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	0.47	9	16	23
	Reading Comprehension	5	1.22	1	3	4
EGRA Ghana English Grade 2	Oral Reading Fluency	60	0.41	39	49	54
	Reading Comprehension	5	1.13	0	3	4

In the above table of benchmarks for 60% RC accuracy, the revised approach incorporates individual subtask standard error estimates (Subtask SE), providing more precise confidence intervals (LCI and UCI) for each subtask. For example, the benchmark for ORF in EGRA Country 4 Arabic Grade 2 is 20, with a narrower confidence interval of 12 to 28, compared to the broader intervals in the initial method. Similarly, in EGRA Ghana English Grade 2, the benchmark for ORF is 49, with a range of 39 to 54, showcasing the tighter bounds enabled by the subtask-specific SE. These narrower intervals indicate a more refined estimate of expected performance, reducing potential overgeneralization. For RC, however, the benchmarks and intervals remain consistent with the initial method, as expected, due to the focus on this skill as the reference point.

Table 10. Rasch Accuracy IRT Benchmarks According to 80% RC and Subtask SE

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.42	19	27	34
	Reading Comprehension	5	1.21	1	4	5
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	0.41	38	46	46
	Reading Comprehension	7	1.1	3	6	6
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	0.56	19	26	29
	Reading Comprehension	5	2.13	1	4	4
EGRA Ghana English Grade 2	Oral Reading Fluency	60	0.5	45	54	54
	Reading Comprehension	5	1.3	1	4	4

The revised benchmarks for 80% RC accuracy also demonstrate the benefits of using subtask-specific SE estimates. For example, in EGRA Country 6 Arabic Grade 2, the benchmark for ORF is 46, with a confidence interval of 38 to 46, compared to the broader intervals in the initial approach. Similarly, in EGRA Country 13 Chitonga Grade 2, the ORF benchmark is 26, with a confidence interval of 19 to 29, reflecting increased precision. This tighter interval suggests greater alignment with the expected ability levels. Across all assessments, the RC benchmarks remain the same as in the initial method, indicating consistency in the measurement of this anchor skill. Overall, the revised approach improves the precision of subtask benchmarks by



accounting for variability across individual subtasks, resulting in a more nuanced and accurate representation of expected performance at the 80% RC proficiency level.

The reduction in SEs in the revised approach is attributable to the conceptual alignment of the benchmark estimation process with subtask-specific variability. Unlike the projection approach, which relies on a singular SE derived from the TIC of RC, the revised approach calculates SEs at the level of each subtask. This method explicitly accounts for the measurement precision of individual subtasks, as determined by the response patterns and item parameters associated with each precursor skill. By focusing on subtask-specific error variances, we achieve a more granular and accurate representation of uncertainty, as opposed to applying the broader, general SE derived from RC to all subtasks.

The substantial reduction in SEs, such as the case for ORF in EGRA Country 13 (from 2.13 to 0.56), stems from the more localized error estimates of the precursor tasks. The initial projection approach applied the SE of RC theta, effectively integrating noise across all subtasks, which inflates the error estimates for individual benchmarks. The revised method isolates the unique measurement error of each subtask, ensuring that the confidence intervals reflect the precision of the specific skill rather than the broader uncertainty of RC. This refined methodology is grounded in psychometric theory, particularly the notion that SEs should directly correspond to the variability inherent in the measurement of the target construct. Thus, the revised approach enhances the validity of the benchmarks by aligning the SEs with the actual measurement context, avoiding the artificial inflation of error margins and providing a clearer, more precise assessment framework.


For example, looking at the case of EGRA Country 13: the SE for OR was 2.13 using the initial “projection” approach, is now only 0.56: a reduction by $\frac{3}{4}$.

Rasch Accuracy - Generalized Linear Mixed Model (RA-GLMM) Approach

To address the issues of large variability and confidence intervals associated with traditional fluency measures, Kara et al. (2020) proposed a novel latent-variable Bayesian model that jointly estimates ORF by simultaneously assessing accuracy and speed. Their model builds on the work of Potgieter et al. (2017), which in turn is an adaptation of the speed-accuracy model introduced by van der Linden (2007). The model proposed by Kara et al. (2020) expands on Potgieter et al.’s (2017) approach by integrating a binomial count factor model for accuracy and a lognormal factor model for speed that are combined through Bayesian estimation. This study expands on the work of Kara et al. (2020) by presenting a modified version of the accuracy component, which will be referred to as the Rasch Accuracy - Generalized Linear Mixed Model (RA-GLMM).

Generalized Linear Mixed Model (GLMM) Approach

In this study, a frequentist approach was employed over a Bayesian framework to analyze ORF, primarily due to the computational efficiency, simplicity, and suitability for large



datasets. Frequentist methods, such as generalized linear mixed models (GLMMs), use optimization-based algorithms like maximum likelihood estimation (MLE), which scale well with large datasets and reduce computational time. Unlike Bayesian methods, which require prior specification and rely on computationally intensive methods of estimation, frequentist approaches avoid the subjectivity of prior selection and focus solely on observed data. As a result, it was determined that a frequentist approach would provide a more transparent, scalable, and efficient method for model estimation, due to the large sample sizes providing sufficient information for parameter estimation.

GLMMs have been widely used in explanatory item analyses to examine how various factors influence student performance and ability estimation (Crocker & Algina, 1986; Kutner, Nachtsheim, & Neter, 2004; van der Linden & Hambleton, 1997). Explanatory item analysis extends traditional item response modeling by incorporating predictors that explain variations in item difficulty, discrimination, and student ability. Unlike standard item response models that estimate item and person parameters as fixed values, explanatory item response models (EIRMs) integrate external variables, such as instructional interventions or cognitive processing strategies, to provide deeper insights into test performance. GLMMs are particularly useful for explanatory item analysis because they allow for the simultaneous estimation of fixed effects (e.g., overall test difficulty) and random effects (e.g., student ability, item variation). By modeling both person-level and item-level influences, GLMMs offer a flexible framework for understanding the factors that contribute to student performance beyond raw test scores.

In this study, a GLMM approach was chosen because the distribution of total test scores and subtask scores did not have normal distributions, making standard linear models unsuitable. GLMMs provided a way to model non-normal response distributions while accounting for hierarchical dependencies in the data, such as students nested within schools or subtasks nested within assessments. Wilson et al. (2012) demonstrated how incorporating IRT into a GLMM framework allowed for the simultaneous estimation of student growth trajectories and item characteristics, illustrating the value of such models for explanatory item analysis. Similarly, Greenwood and Jesse (2014) applied GLMMs to analyze binary item responses in a longitudinal study, demonstrating how the approach allowed for a more precise estimation of ability changes over time. In this study, for consistency with the ORF model proposed by Kara et al. (2020) and to maintain computational efficiency, a binomial distribution was applied at the subtask level rather than modeling each item individually.

RA-GLMM Estimation

The RA-GLMM is much like the accuracy component of the latent ORF ability model proposed by Kara et al. (2020), in that it models the number of correct responses for a given subtask. However, in the model proposed by Kara et al. (2020) this is relegated only to ORF, in the

present study this is generalized to all subtasks. Expected accuracy is calculated using the following equations:

$$E[U_{ij}] = n_i \cdot p(U_{ij}),$$

$$p(U_{ij}) = \frac{e^{1.7(\theta_{ij}-b_i)}}{1+e^{1.7(\theta_{ij}-b_i)}},$$

where:

- b_i is the difficulty parameter of subtask i ,
- θ_{ij} is the latent accuracy ability of person j on subtask i ,
- 1.7 is a scaling constant value which adjusts the logistic function to approximate the normal ogive model by accounting for the difference in their variances with $1.7 \approx \pi/\sqrt{3}$ to ensure comparable item difficulty estimates (Baker & Kim, 2004),
- U_{ij} is a proportion of correct responses on subtask i for person j .
- and n_i is the total possible score on subtask i .

The expression $e^{1.7(\theta_{ij}-b_i)}$ represents the odds of achieving a perfect score on the subtask based on the student's ability and the subtask difficulty, and $\frac{e^{1.7(\theta_{ij}-b_i)}}{1+e^{1.7(\theta_{ij}-b_i)}}$ converts the odds into a probability (via logistic transformation). Multiplying this probability by n_i provides the expected score the student will achieve on the subtask. An explanation of the process is as follows:

- θ_{ij} is extracted from a mixed-effects logistic regression model that uses the observed total test score as the response variable and includes a random effect for each student.
 - The model predicts the probability of a correct response proportion for the entire test:

$$\text{logit}(p_{test}) = \theta_j + u_j,$$

- Where:
 - p_{test} is the probability of a proportion of correct responses for student j on the entire assessment (including all subtasks).
 - θ_j is the latent ability of student j (fixed effect).
 - u_j is the random effect for student j , representing unexplained variance.
- b_i is derived from a logistic regression model that uses theta as the predictor for observed scores.
- The logistic transformation ensures that the probability of subtask score $p(U_{ij})$ lies between 0 and 1.

- The expected probability $p(U_{ij})$ is multiplied by the subtask's maximum score n_i to calculate the expected subtask score.

Interpretively, the accuracy component of the RA-GLMM is as follows:

- As θ_{ij} increases (indicating higher student ability), the expected accuracy ($E[U_{ij}]$) increases, reflecting a higher likelihood of increased accuracy.
- Passages with higher difficulty b_i lower the expected accuracy for all students, as the logistic function shifts higher.
- The passage length determines the upper bound of the expected accuracy, ensuring consistency across passages of varying lengths.

RA-GLMM TCC

The total test characteristic curve (TCC) is the sum of the expected scores for all subtasks:

$$TCC(\theta) = \sum_i n_i \cdot p(U_i),$$

The standard error at a given θ is derived from the variance of the TCC. Since the subtasks are assumed to be independent, the variance of the total score is the sum of the variances of the scores for each subtask. The SE for the total score is then:

$$SE(\theta_{test}) = \sqrt{\sum_i \text{Var}(U_i)}.$$

However, to calculate the SE of θ for each subtask i , the information function of the logistic model must be employed (like the IRT approach). The Fisher Information ($I(\theta)$) for a subtask is given by:

$$I(\theta) = 1.7^2 \cdot n_i^2 \cdot p(U_i) \cdot (1 - p(U_i)),$$

Where:

- $I(\theta)$ is the Fisher information at a given θ .
- $p(U_i)$ is the probability of a proportion of correct responses on subtask i .
- n_i is the maximum score for the subtask.


Thus, the standard error is the inverse of the square root of the Fisher information:

$$SE(\theta_i) = \frac{1}{\sqrt{I(\theta)}}.$$

Consequently, the confidence intervals for the RA-GLMM accuracy component can be calculated by $\theta_{RC} \pm 1.96 \times SE(\theta_i)$, closely resembling the IRT method.

Summary of RA-GLMM and Attributes

The RA-GLMM was developed to provide a more stable and interpretable approach to benchmarking student performance on subtasks, particularly in foundational literacy



assessments. Traditional fluency measures often exhibit large variability and wide confidence intervals, making it difficult to set consistent benchmarks. To address this, Kara et al. (2020) introduced a latent-variable Bayesian model that jointly estimates accuracy and speed. This study builds on this approach by modifying the accuracy component and applying a GLMM framework instead of a Bayesian estimation approach. The GLMM approach was chosen for its computational efficiency and scalability, especially when working with large datasets where Bayesian methods may be impractical. By using a binomial distribution for subtask-level scores rather than modeling each item individually, the RA-GLMM simplifies parameter estimation while maintaining alignment with the Rasch modeling framework.

The choice to use a GLMM was also driven by the fact that the distributions of total test scores and subtask scores did not closely follow a normal distribution, making standard linear models unsuitable. GLMMs allow for the modeling of non-normal response distributions while accounting for hierarchical dependencies, such as subtasks nested within assessments. The RA-GLMM takes advantage of these strengths by estimating expected accuracy at the subtask level, ensuring that benchmark-setting is both statistically sound and practically interpretable.

RA-GLMM Results

The data were analyzed using the R software package lme4 (Bates et al., 2015), with the bound optimization by quadratic approximation (BOBYQA) optimizer and a maximum of 100,000 function evaluations to ensure convergence, which employs a frequentist mixed-effects modeling approach through MLE to fit GLMMs. The BOBYQA optimizer is a derivative-free optimization method that is generally better suited for handling convergence issues in mixed models (Powell, 2009). This method allows for the inclusion of both fixed and random effects, enabling the accurate modeling of hierarchical or clustered data structures, such as repeated measures or student-level variability. Students that read at least one correctly and used less than 10 seconds of time were excluded from the analysis as it was determined that at least 10 seconds would be considered sufficient to provide reliable estimates.

Table 10. Rasch Accuracy GLMM Benchmarks According to 60% RC and Subtask SE

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.03	21	22	23
	Reading Comprehension	5	0.24	2	3	4
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	0.02	29	29	30
	Reading Comprehension	7	0.17	3	4	5
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	0.03	9	9	10
	Reading Comprehension	5	0.24	2	3	4
EGRA Ghana English Grade 2	Oral Reading Fluency	60	0.02	46	47	48
	Reading Comprehension	5	0.24	2	3	4

The table above presents benchmarks derived from the RA-GLMM. The key differences between these results and those of the Rasch Accuracy IRT method are reflected in the benchmark values, LCI, and UCI for each subtask and assessment. For ORF, the GLMM-derived benchmarks tend to have narrower confidence intervals (e.g., EGRA Country 4 Arabic Grade 2: LCI = 21, UCI = 23) compared to the IRT-derived benchmarks (LCI = 12, UCI = 28). This suggests that the GLMM approach may produce more precise estimates for the benchmarks. Similarly, the GLMM benchmarks for RC are consistent, with minimal variability across assessments (e.g., EGRA Country 6 Arabic Grade 2: Benchmark = 4, UCI = 5), whereas the IRT model exhibits broader intervals and variability (e.g., the same assessment: Benchmark = 4, UCI = 6).

The θ_{SE} column highlights notable differences in standard errors between the two models. The GLMM approach consistently produces smaller standard errors (e.g., EGRA Country 6 Arabic Grade 2 ORF: 0.02) than the IRT model (e.g., same assessment ORF: 0.45). Overall, the GLMM benchmarks display tighter ranges and reduced variability compared to the IRT benchmarks, suggesting that the GLMM method may offer greater precision and consistency. However, the broader intervals in the IRT model could indicate a more conservative estimation approach, capturing greater uncertainty in the data. These differences highlight the importance of selecting an appropriate modeling technique based on the assessment goals and data characteristics.

Table 11. Rasch Accuracy GLMM Benchmarks According to 80% RC and Subtask SE

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
EGRA Country 4 Arabic Grade 2	Oral Reading Fluency	42	0.03	31	32	32
	Reading Comprehension	5	2.4	3	4	5
EGRA Country 6 Arabic Grade 2	Oral Reading Fluency	76	0.02	55	56	57
	Reading Comprehension	7	2.15	5	6	7
EGRA Country 13 Chitonga Grade 2	Oral Reading Fluency	56	0.02	19	19	20
	Reading Comprehension	5	1.88	3	4	5
EGRA Ghana English Grade 2	Oral Reading Fluency	60	0.03	54	54	55
	Reading Comprehension	5	2.51	3	4	5

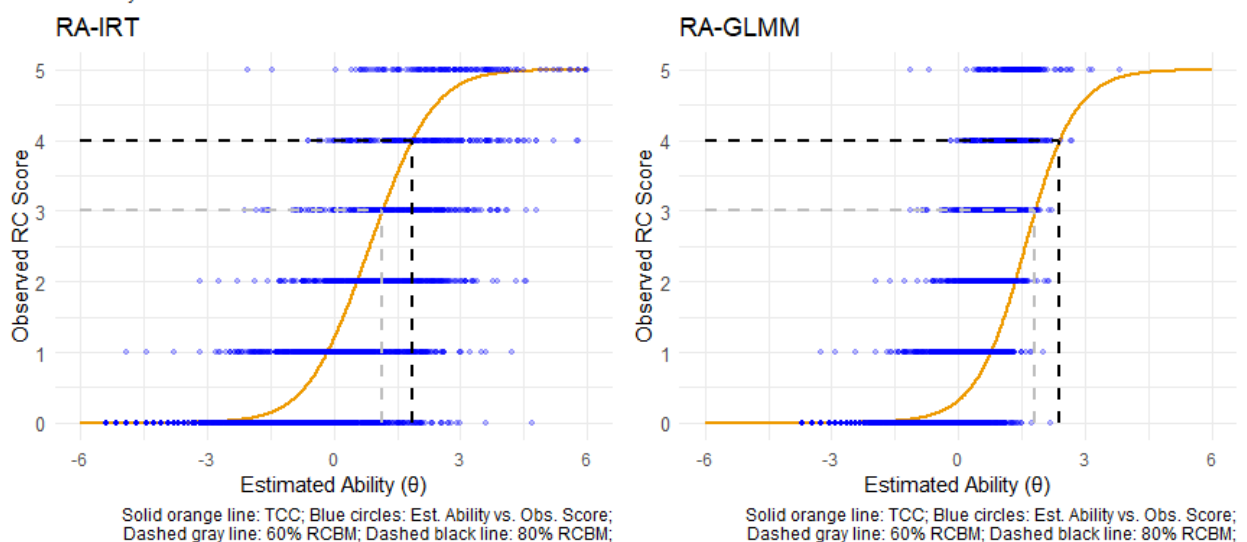
The table above compares benchmarks derived from GLMM for 80% RC. For ORF, GLMM consistently produces higher benchmark values with narrower CI than Rasch Accuracy IRT estimates, indicating greater precision. For example, in EGRA Country 6 Arabic Grade 2, the GLMM benchmark is 56 with a CI range of 55–57, compared to the IRT benchmark of 46 with a broader CI range of 38–46. Similar trends are observed in other assessments, such as EGRA Country 4 Arabic Grade 2, where GLMM produces a benchmark of 32 with a tight CI range of 31–32, while IRT yields a lower benchmark of 27 and a wider CI range of 19–34.

For RC, GLMM also demonstrates more consistent benchmarks across assessments with relatively narrower confidence intervals. For instance, in EGRA Ghana English Grade 2, the GLMM benchmark is 4 with a CI range of 3–5, while the IRT benchmark shows greater variability with a CI range of 1–4. Although SE for RC are slightly higher in GLMM compared to IRT (e.g., 2.4 for EGRA Country 4 Arabic Grade 2 in GLMM versus 1.21 in IRT), GLMM still maintains better overall precision.

The narrower CI widths in GLMM highlight its reliability, particularly for ORF, where the SEs are significantly smaller (e.g., 0.02–0.03) compared to IRT (e.g., 0.41–0.56). For RC, while both models display wider CIs, GLMM generally provides tighter ranges, offering more reliable benchmarks. In contrast, IRT’s broader CIs and lower benchmarks reflect a more conservative approach, which captures greater variability but sacrifices precision.

Figure 8. Reading Comprehension (RC) TCC’s: Country 4 Arabic Grade 2

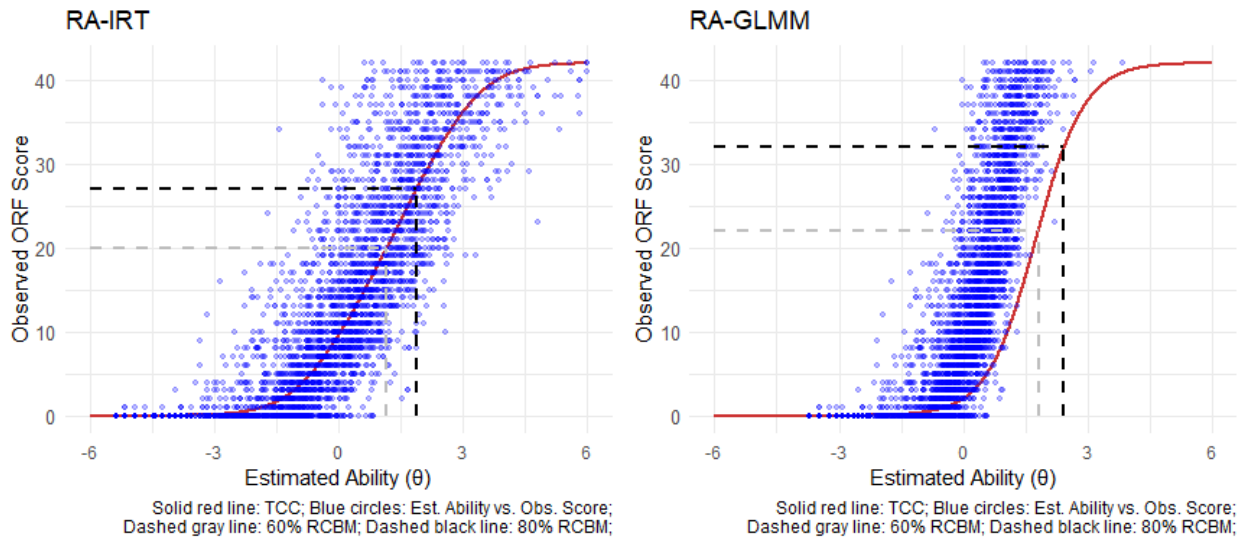
EGRA Country 4 Arabic Grade 2



The figure above compares ability estimates and observed reading comprehension scores using the RA-IRT and RA-GLMM methods. In the RA-IRT model, ability estimates are more widely distributed, meaning there is greater variability, but they align closely with the test characteristic curve. In contrast, the RA-GLMM model produces a more compressed range of estimates, resulting in tighter standard errors and greater precision.

Figure 9. Oral Reading Fluency (ORF) TCC's: Country 4 Arabic Grade 2

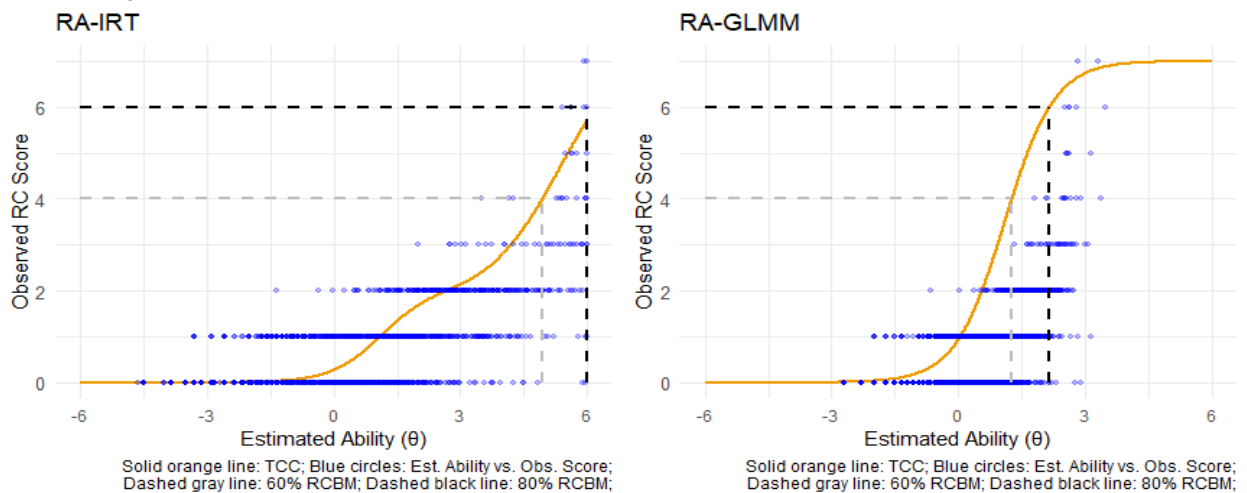
EGRA Country 4 Arabic Grade 2



The figure above compares estimated ability and observed ORF scores using the RA-IRT and RA-GLMM methods. The RA-IRT model shows a wider spread of ability estimates, allowing for greater variability across students while still aligning closely with the test characteristic curve. In contrast, the RA-GLMM model produces a more compressed distribution of ability estimates, leading to tighter standard errors and more precise predictions. Despite these differences, both methods yield similar benchmark placements for 60% and 80% RC accuracy.

Figure 10. Reading Comprehension (RC) TCC's: Country 6 Arabic Grade 2

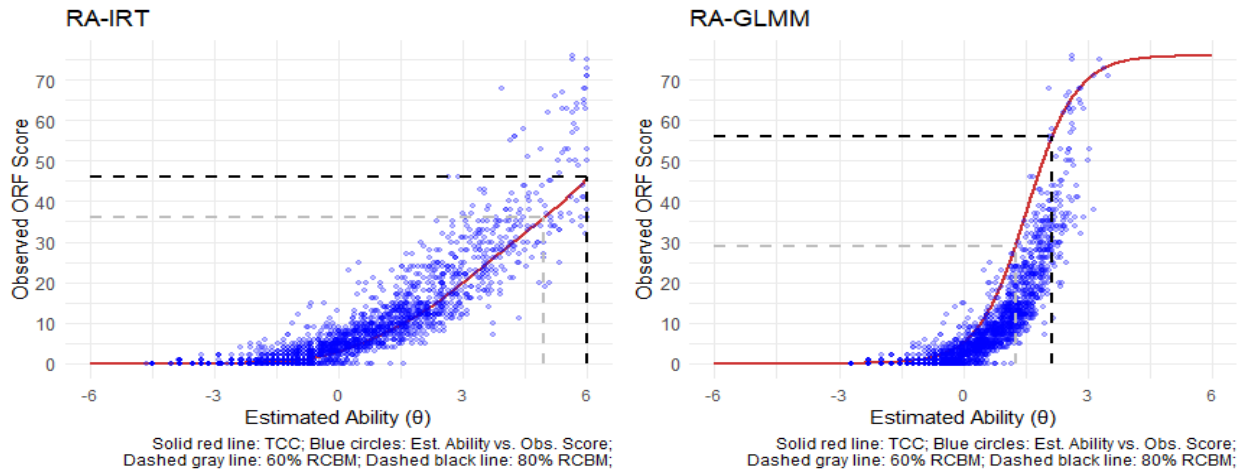
EGRA Country 6 Arabic Grade 2



The figure above compares estimated ability and observed RC scores for RA-IRT and RA-GLMM methods in EGRA Country 6 Arabic Grade 2. The RA-IRT model produces a wider spread of ability estimates, capturing more variability across students while still aligning with the test characteristic curve. The RA-GLMM model, on the other hand, constrains ability estimates more tightly, leading to lower standard errors and greater precision.

Figure 11. Oral Reading Fluency (ORF) TCC's: Country 6 Arabic Grade 2

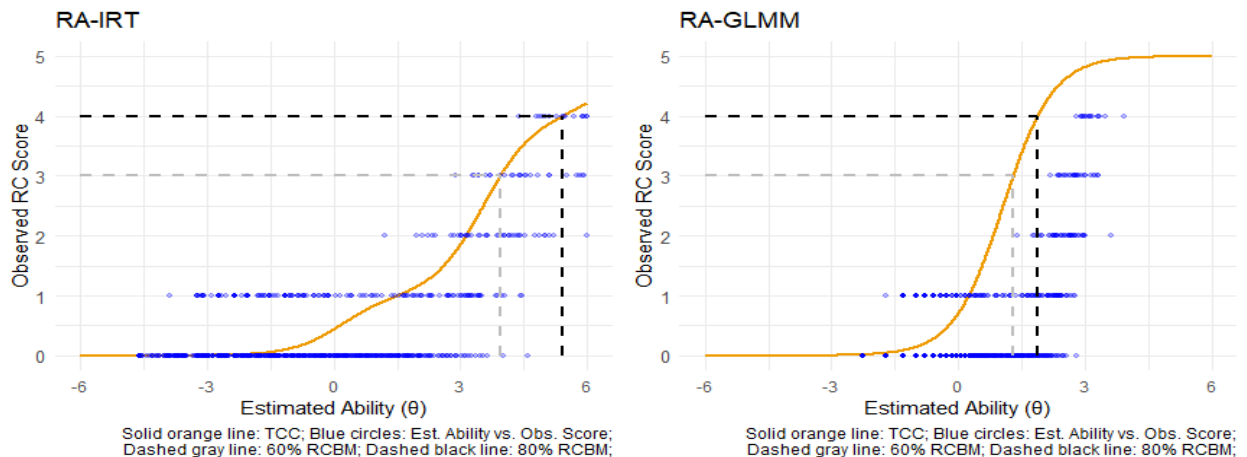
EGRA Country 6 Arabic Grade 2



The figure above compares estimated ability and observed ORF scores for RA-IRT and RA-GLMM methods in EGRA Country 6 Arabic Grade 2. The RA-IRT model shows a wider spread of ability estimates, capturing greater variability in student performance while aligning closely with the test characteristic curve. In contrast, the RA-GLMM model produces a more constrained distribution of ability estimates, resulting in lower standard errors and more precise predictions. This suggests that while IRT allows for a broader range of ability levels, GLMM stabilizes estimates by reducing variability. Despite these differences, both models place the 60% and 80% RC benchmarks in similar locations.

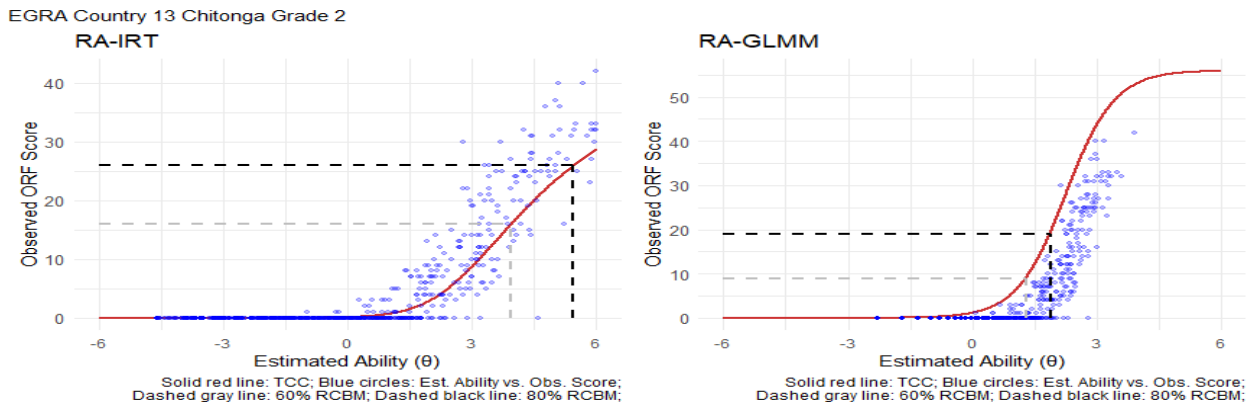
Figure 12. Reading Comprehension (RC) TCC's: Country 13 Chitonga Grade 2

EGRA Country 13 Chitonga Grade 2



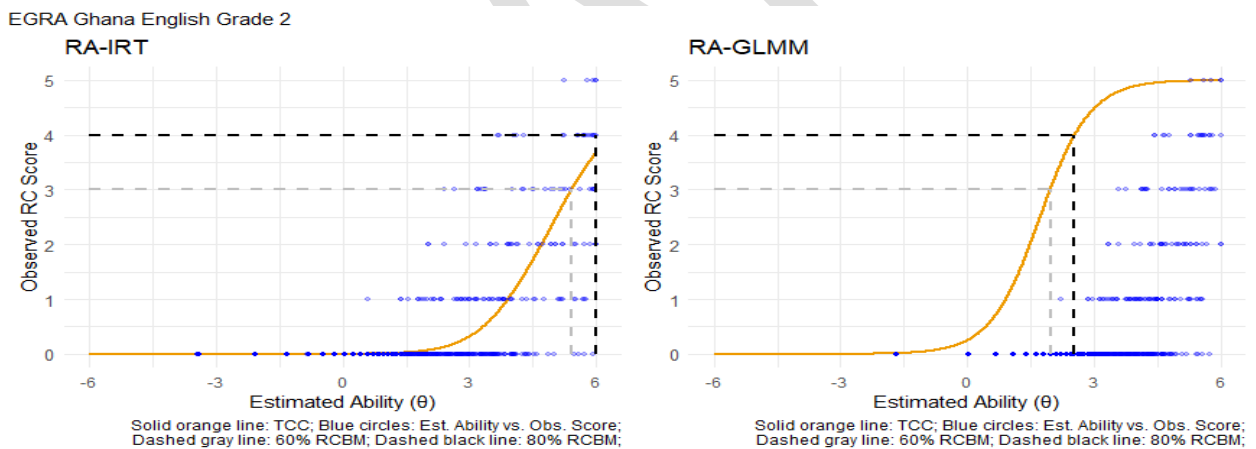
The figure above compares estimated ability and observed RC scores for RA-IRT and RA-GLMM methods in EGRA Country 13 Chitonga Grade 2. The RA-IRT model displays a wider distribution of ability estimates, capturing greater variability among students while closely following the test characteristic curve. The RA-GLMM model, in contrast, constrains ability estimates more tightly, resulting in lower SEs and more precise predictions.

Figure 13. Oral Reading Fluency (ORF) TCC's: Country 13 Chitonga Grade 2



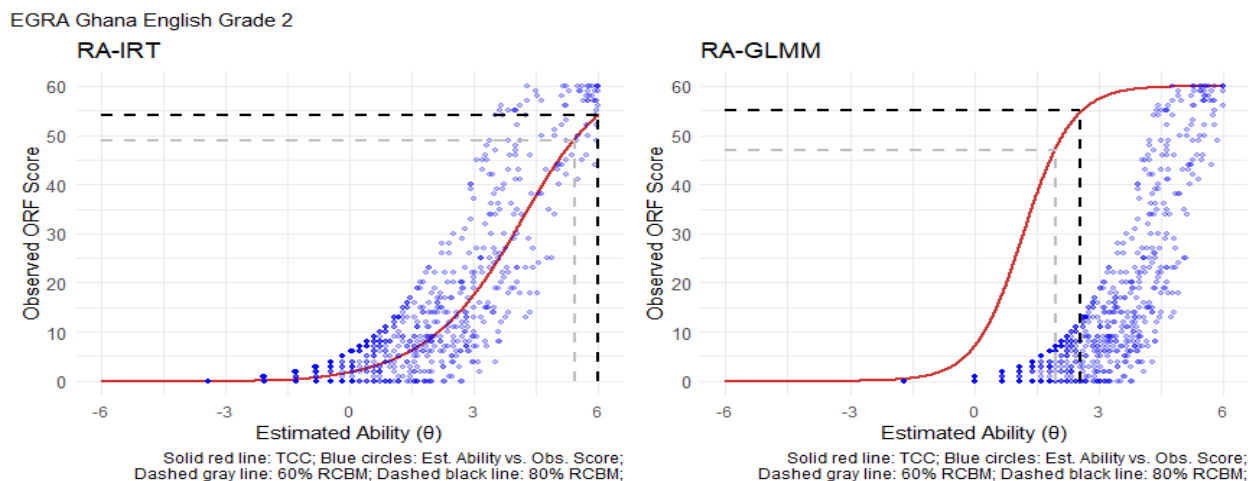
The figure above compares estimated ability and observed ORF scores for RA-IRT and RA-GLMM methods in EGRA Country 13 Chitonga Grade 2. The RA-IRT model shows a wider spread of ability estimates, capturing more variability among students while aligning closely with the test characteristic curve. In contrast, the RA-GLMM model constrains ability estimates more tightly, resulting in lower SEs and greater precision. Despite these differences, both models place the 60% and 80% RC benchmarks in similar locations.

Figure 14. Reading Comprehension (RC) TCC's: Ghana English Grade 2



The figure above compares estimated ability and observed RC scores for RA-IRT and RA-GLMM methods in EGRA Ghana English Grade 2. The RA-IRT model shows a wider distribution of ability estimates, capturing more variability across students while still aligning with the test characteristic curve. In contrast, the RA-GLMM model constrains ability estimates more tightly, leading to lower SEs and more precise predictions.

Figure 15. Oral Reading Fluency (ORF) TCC's: Ghana English Grade 2



The figure above compares estimated ability and observed ORF scores for RA-IRT and RA-GLMM methods in EGRA Ghana English Grade 2. The RA-IRT model shows a wider spread of ability estimates, capturing greater variability among students while aligning closely with the test characteristic curve. In contrast, the RA-GLMM model produces a more constrained distribution of ability estimates, leading to lower SEs and more precise predictions. Despite these differences, both models place the 60% and 80% reading comprehension benchmarks in similar locations.

Method Selection for Benchmarking

Although the GLMM approach demonstrated superior performance compared to the Rasch IRT and classical methods in terms of the results obtained, it is not without its limitations. The RA-GLMM approach, while flexible and capable of modeling complex data structures, has significant limitations when applied to benchmark estimation in educational assessments. One major drawback is its inability to account for item-level differences explicitly. In a GLMM framework, the primary focus is on aggregating responses across individuals and items without directly modeling the unique characteristics of each item. This means that variations in item difficulty, discrimination, or other item parameters are not incorporated into the model. As a result, GLMM tends to treat all items as if they contribute equally to the estimation of ability, which can lead to biased or imprecise results, particularly in assessments where item-level differences are substantial. By ignoring these nuances, GLMM may fail to capture critical aspects of the data that influence the accuracy of ability estimates.

In contrast, the Rasch IRT model offers a more robust framework by calibrating all items simultaneously and placing them on a common ability scale. This simultaneous calibration ensures that item-level differences, such as varying levels of difficulty, are explicitly accounted for in the estimation process. By anchoring the ability scale to the characteristics of the items, the Rasch approach provides a unified and interpretable measurement system where the estimated ability of an individual is directly comparable across all items in the assessment. This is particularly advantageous in educational contexts where tests may consist of items

with differing levels of complexity or partial credit scoring. The Rasch model's ability to place items and individuals on the same scale facilitates more accurate benchmarking and ensures that the interpretation of scores is consistent across varying test forms or populations.

Another critical advantage of the Rasch approach is its capacity to provide invariant measurement, which ensures that the ability estimates are independent of the specific items used and vice versa. This invariance is not guaranteed in GLMM because the model does not explicitly calibrate items or link them to a shared scale. Furthermore, Rasch's focus on calibrating all items simultaneously enables a deeper understanding of item functioning and test performance, including the detection of misfitting items or differential item functioning (DIF) across subpopulations. These insights are crucial for ensuring the fairness and validity of assessments. In contrast, GLMM's reliance on aggregated measures can obscure such issues, potentially leading to misleading conclusions about student ability or test quality. Ultimately, while GLMM may be useful for simpler applications, the Rasch approach's explicit attention to item-level differences and its simultaneous calibration of items make it a superior choice for rigorous, fair, and interpretable benchmarking in educational assessments.

Section 5: Data Analysis of Additional Countries and Languages Using Accuracy-Based Model

After assessing the suitability of various methods using a limited set of four test cases consisting of language-country pairing, the study shifted focus to benchmarking analysis for the broader range of languages in the datasets provided by UIS. Our approach began with the establishment of basic reliability parameters for all the language-country combinations, followed by the benchmark estimation process.

Table 12. EGRA – Reliability by Assessment

Assessment	N	Alpha
EGRA Country 4 Arabic Grade 2	615	0.988
EGRA Country 6 Arabic Grade 2	713	0.986
EGRA Country 7 Chichewa Grade 2	713	0.987
EGRA Country 10 Kinyarwanda Grade 2	615	0.993
EGRA Country 11 Swahili Grade 2	713	0.990
EGRA Country 13 Chitonga Grade 2	615	0.988
EGRA Country 13 Cinyanja Grade 2	615	0.988
EGRA Country 13 Icibemba Grade 2	615	0.987
EGRA Country 13 Kikaonde Grade 2	615	0.987
EGRA Country 13 Lunda Grade 2	615	0.989
EGRA Country 13 Luvale Grade 2	615	0.989
EGRA Country 13 Silozi Grade 2	615	0.988
EGRA Ghana English Grade 2	713	0.988

The table above presents the number of observations (N) and reliability estimates (Cronbach's Alpha) for various Grade 2 assessments across different countries and languages. The data reflects high reliability across all assessments, with Cronbach's Alpha values ranging from 0.986 to 0.993. These values indicate excellent internal consistency, suggesting that the items within each assessment are highly correlated and measure the same underlying construct effectively. Columns with no variance were removed prior to generating this summary.

Among the assessments, EGRA Country 10 Kinyarwanda Grade 2 stands out with the highest reliability estimate (Alpha = 0.993). Other assessments, such as those for EGRA Country 6 Arabic Grade 2 (Alpha = 0.986) and EGRA Country 13 Icibemba Grade 2 (Alpha = 0.987), also show strong reliability, though slightly lower. The consistent high Alpha values across assessments and languages underscore the robustness of the test instruments used, reflecting their suitability for measuring student abilities across diverse contexts.

Table 13. EGRA – RC and ORF Item-Test Correlation by Assessment

Assessment	RC1	RC2	RC3	RC4	RC5	RC6	RC7	ORF
Country 4 Arabic	0.41	0.32	0.06	0.26	0.25			0.77
Country 6 Arabic	0.47	0.42	0.19	0.34	0.01	0.35	0.3	0.91
Country 7 Chichewa	0.74	0.78						0.66
Country 10 Kinyarwanda	0.11	0.14	0.12	0.1	0.17			0.44
Country 11 Swahili	0.33	0.33	0.21	0.16	0.46			0.96
Country 13 Chitonga	0.48	0.61	0.46	-0.03	-0.27			0.89
Country 13 Cinyanja	0.71	0.53	0.42	0.37				0.92
Country 13 Icibemba	0.53	0.41	0.13	0.49	1			0.9
Country 13 Kikaonde	0.61	0.48	0.28	0.34	-0.56			0.87
Country 13 Lunda	0.72	0.6	0.29	0.15	-0.03			0.94
Country 13 Luvale	0.77	0.51	0.57	0.55	-0.88			0.95
Country 13 Silozi	0.43	0.54	0.45	0.65				0.93
Ghana English	0.55	0.6	0.43	0.39	0.25			0.92

The table above presents the item-test correlation coefficients for RC (RC1-RC7) and ORF across various EGRA assessments. The correlations provide insight into the alignment between individual items and the overall test performance for each assessment. Assessments like Country 11 Swahili and Country 13 Luvale exhibit consistently high RC correlations, with values reaching up to 0.77 for RC1 and RC5, and strong ORF correlations of 0.96 and 0.95, respectively. Most assessments show moderate to strong ORF correlations (0.77 to 0.96), with notable exceptions like Country 7 Chichewa (0.66).

Table 14. EGRA - Bartlett's Test of Sphericity

Assessment	Chi-Sq	P	df
Country 4 Arabic Grade 2	17861	<.001	21
Country 6 Arabic Grade 2	9182	<.001	15
Country 7 Chichewa Grade 2	3212	<.001	6

Country 10 Kinyarwanda Grade 2	7316	<.001	15
Country 11 Swahili Grade 2	37606	<.001	6
Country 13 Chitonga Grade 2	5072	<.001	15
Country 13 Cinyanja Grade 2	12215	<.001	15
Country 13 Icibemba Grade 2	6834	<.001	15
Country 13 Kikaonde Grade 2	3591	<.001	15
Country 13 Lunda Grade 2	8571	<.001	15
Country 13 Luvale Grade 2	5805	<.001	15
Country 13 Silozi Grade 2	8463	<.001	15
Ghana English Grade 2	4572	<.001	10

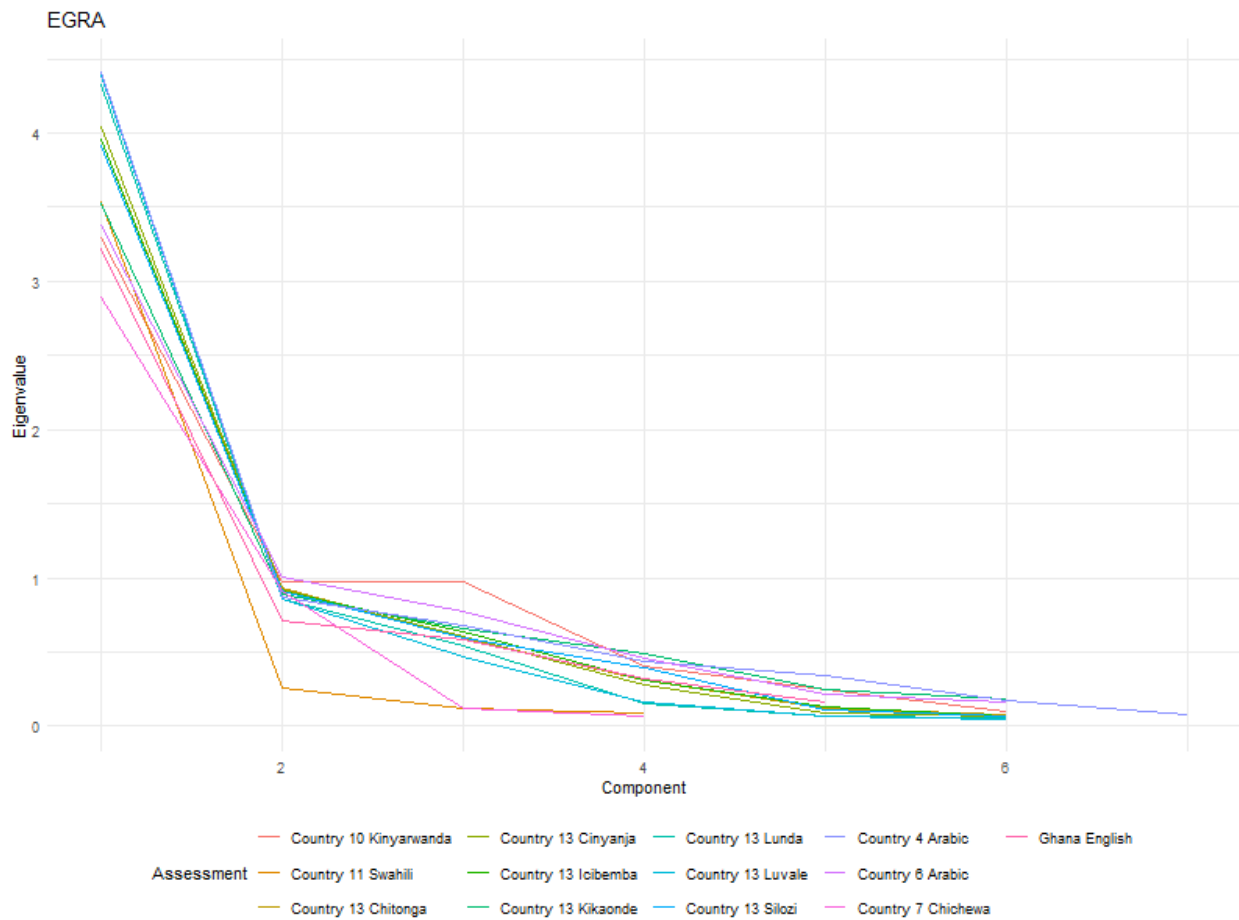
The results of Bartlett’s Test of Sphericity for the EGRA assessments across various languages and countries indicate statistically significant Chi-square values ($p = 0$ for all tests), suggesting that the correlation matrices are appropriate for factor analysis. Notably, the Chi-square values range from 3,212 (Country 7 Chichewa Grade 2) to 37,606 (Country 11 Swahili Grade 2), with degrees of freedom varying between 6 and 21. These findings confirm the suitability of the datasets for exploring underlying latent structures, reinforcing the validity of proceeding with further multivariate analyses for these assessments.

Table 15. EGRA - Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Assessment	Overall	LC	LS	SS	INW	ORF	RC	SRC
Country 4 Arabic	0.81	0.83	0.78	0.88	0.82	0.75	0.73	0.90
Country 6 Arabic	0.81	0.85	0.83		0.83	0.77	0.80	
Country 7 Chichewa	0.63			0.62		0.64	0.64	
Country 10 Kinyarwanda	0.81	0.72		0.78		0.87	0.90	
Country 11 Swahili	0.86			0.87	0.83	0.83	0.91	
Country 13 Chitonga	0.84	0.93	0.93	0.83	0.82	0.80	0.84	
Country 13 Cinyanja	0.85	0.85	0.88	0.83	0.87	0.82	0.89	
Country 13 Icibemba	0.85	0.91	0.90	0.92	0.83	0.78	0.87	
Country 13 Kikaonde	0.83	0.95	0.79	0.79	0.93	0.87	0.81	
Country 13 Lunda	0.88	0.95	0.97	0.89	0.87	0.84	0.88	
Country 13 Luvale	0.87	0.94	0.91	0.89	0.87	0.82	0.89	
Country 13 Silozi	0.86	0.93	0.90	0.87	0.83	0.81	0.94	
Ghana English	0.80	0.89	0.84		0.83	0.74	0.76	

This table presents reliability coefficients for various EGRA across multiple countries and languages, focusing on overall reliability and subtest-level reliability. Overall reliability values range from 0.63 (Country 7 Chichewa) to 0.88 (Country 13 Lunda), demonstrating generally high reliability for most assessments. Subtest-level reliability coefficients also reflect consistency, with LC, LS, SS, INW, ORF, RC, and SRC showing strong internal consistency for most countries. Notable variability exists in lower coefficients for some subtests in Country 7 Chichewa, suggesting areas for further investigation or improvement. Overall, the data highlights the robustness of the assessment tools across diverse linguistic and cultural contexts.

Figure 16. EGRA - Scree Plot



This scree plot depicts the eigenvalues for various components across multiple EGRA assessments, allowing for visual comparison of the explained variance by each principal component. For most assessments, the first component explains most of the variance, as indicated by the sharp drop in eigenvalues from Component 1 to Component 2. The consistent shape across assessments indicates homogeneity in the dimensionality of reading-related constructs measured across countries. The rapid decline of eigenvalues reinforces the appropriateness of focusing on fewer components for subsequent analysis, particularly for reducing dimensionality without significant loss of information.

Table 16. EGRA - Principal Components Analysis (PCA)

Assessment	λ	F1 % Var	L1	L2	L3	L4	L5	L6	L7
Country 4 Arabic	4.41	0.63	0.48	0.64	0.88	0.87	0.92	0.85	0.81
Country 6 Arabic	3.38	0.56	0.45			0.85	0.91	0.83	
Country 7 Chichewa	2.90	0.72							
Country 10 Kinyarwanda	3.29	0.55	0.23		0.92		0.84	0.87	
Country 11 Swahili	3.54	0.88							
Country 13 Chitonga	3.96	0.66	0.32	0.73	0.92	0.94	0.93	0.85	
Country 13 Cinyanja	4.04	0.67	0.34	0.73	0.95	0.94	0.95	0.84	

Country 13 Icibemba	3.95	0.66	0.40	0.65	0.87	0.95	0.94	0.91	
Country 13 Kikaonde	3.53	0.59	0.40	0.71	0.91	0.75	0.85	0.87	
Country 13 Lunda	4.33	0.72	0.44	0.72	0.95	0.96	0.96	0.94	
Country 13 Luvale	4.39	0.73	0.45	0.78	0.96	0.95	0.96	0.92	
Country 13 Silozi	3.92	0.65	0.34	0.70	0.94	0.94	0.94	0.80	
Ghana English	3.22	0.64	0.63	0.76					

This table presents the results of PCA for the EGRA assessments, summarizing the first factor's explained variance percentage (F1 % Var) and loadings across multiple components (L1 to L7) for each assessment. The explained variance by the first factor (F1 % Var) varies across assessments, ranging from 2.90 in Country 7 Chichewa to 4.41 in Country 4 Arabic, indicating that the first factor captures a substantial proportion of the variance in the data. Assessments such as Country 13 Luvale and Country 13 Lunda exhibit high explained variance (4.39 and 4.33, respectively), suggesting a strong single-dimension structure.

Table 17. EGRA - Accuracy Benchmarks According to 60% RC and Subtask SE

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
Country 4 Arabic	Oral Reading	1	0.42	12	20	28
	Reading Comprehension	5	1.03	0	3	5
Country 6 Arabic	Oral Reading	1	0.45	29	36	44
	Reading Comprehension	7	1.04	2	4	6
Country 7 Chichewa	Oral Reading	1	0.45	11	17	17
	Reading Comprehension	4	5.24	0	2	2
Country 10 Kinyarwanda	Oral Reading	1	0.58	26	37	39
	Reading Comprehension	5	1.05	1	3	5
Country 11 Swahili	Oral Reading	1	0.53	17	24	30
	Reading Comprehension	5	1.25	0	3	5
Country 13 Chitonga	Oral Reading	1	0.47	9	16	23
	Reading Comprehension	5	1.22	1	3	4
Country 13 Cinyanja	Oral Reading	1	0.57	17	23	29
	Reading Comprehension	5	1.21	0	3	4
Country 13 Icibemba	Oral Reading	1	0.52	6	12	18
	Reading Comprehension	5	1.48	0	3	4
Country 13 Kikaonde	Oral Reading	1	0.47	11	18	25
	Reading Comprehension	5	1.57	0	3	4
Country 13 Lunda	Oral Reading	1	0.53	8	14	20
	Reading Comprehension	5	1.4	1	3	5
Country 13 Luvale	Oral Reading	1	0.58	16	23	29
	Reading Comprehension	5	1.35	1	3	4
Country 13 Silozi	Oral Reading	1	0.54	21	27	33
	Reading Comprehension	5	1.73	1	3	4
Ghana English	Oral Reading	1	0.41	39	49	54
	Reading Comprehension	5	1.13	0	3	4

The table above highlights the variation in ORF benchmarks across different assessments. These benchmarks, based on achieving a 60% RC threshold, show significant differences between languages and contexts. For example, Ghana English exhibits the highest benchmark at 49 words, with a CI ranging from 39 to 54 words, indicating higher expectations for reading fluency. In contrast, assessments such as Country 13 Ibibemba and Country 13 Chitonga have lower benchmarks, at 12 and 16 words respectively, with narrower CI. Countries like Country 6 Arabic and Country 10 Kinyarwanda also display relatively high benchmarks at 36 and 37 words, suggesting a higher standard of oral fluency compared to other contexts. Conversely, several assessments, including Country 7 Chichewa and Country 13 Kikaonde, show moderate benchmarks of 17 and 18 words, reflecting more conservative expectations for fluency.

Table 18. EGRA - Benchmarks According to 80% RC and Subtask SE

Assessment	Subtask	Item	θ_{SE}	LCI	Benchmark	UCI
Country 4 Arabic	Oral Reading	1	0.42	19	27	34
	Reading Comprehension	5	1.21	1	4	5
Country 6 Arabic	Oral Reading	1	0.41	38	46	46
	Reading Comprehension	7	1.1	3	6	6
Country 7 Chichewa	Oral Reading	1	0.45	11	17	17
	Reading Comprehension	4	5.24	0	3	3
Country 10 Kinyarwanda	Oral Reading	1	0.96	26	39	40
	Reading Comprehension	5	1.21	1	4	5
Country 11 Swahili	Oral Reading	1	0.59	23	29	35
	Reading Comprehension	5	1.39	1	4	5
Country 13 Chitonga	Oral Reading	1	0.56	19	26	29
	Reading Comprehension	5	2.13	1	4	4
Country 13 Cinyanja	Oral Reading	1	0.57	29	34	34
	Reading Comprehension	5	5.2	0	4	4
Country 13 Ibibemba	Oral Reading	1	0.54	17	23	29
	Reading Comprehension	5	2.26	1	4	4
Country 13 Kikaonde	Oral Reading	1	0.45	21	29	31
	Reading Comprehension	5	1.46	2	4	4
Country 13 Lunda	Oral Reading	1	0.56	14	21	26
	Reading Comprehension	5	1.45	1	4	5
Country 13 Luvale	Oral Reading	1	0.56	24	30	30
	Reading Comprehension	5	1.74	1	4	4
Country 13 Silozi	Oral Reading	1	0.59	27	33	33
	Reading Comprehension	5	1.93	1	4	4
Ghana English	Oral Reading	1	0.5	45	54	54
	Reading Comprehension	5	1.3	1	4	4

The table above highlights the benchmarks for ORF derived from an 80% RC threshold, emphasizing variability across assessments. Ghana English displays the highest ORF benchmark at 54 correct words, with a CI ranging from 45 to 54, reflecting the highest fluency

expectations among the assessments. Country 6 Arabic follows with a benchmark of 46 correct words (CI: 38–46), while other assessments such as Country 10 Kinyarwanda and Country 13 Cinyanja exhibit moderate benchmarks of 39 and 34 correct words, respectively.

In contrast, some assessments have lower ORF benchmarks, such as Country 13 Lunda (21 correct words, CI: 14–26) and Country 13 Chitonga (26 correct words, CI: 19–29). Country 7 Chichewa displays the lowest benchmark at 17 correct words (CI: 11–17), highlighting varied expectations for reading fluency across languages and contexts. These differences underscore the diverse linguistic and educational environments influencing reading performance and benchmarks.

Section 6: Conclusions and Summary

The primary objective of this study has been to integrate recently developed foundational literacy assessments—which are typically one-on-one, oral, and have conceptually complex test “item” frameworks—into a more rigorous and comparably psychometric framework for analysis. The significance of this task extends beyond a technical or academic exercise—it is crucial for linking assessments to globally agreed MPLs and fulfilling the SDG mandate for comparability, even in cases where assessments were not originally designed with such comparability.

In summary, the findings from Sections 3 and 4 confirm that IRT is a reliable and effective method for setting benchmarks in literacy assessments, particularly for RC and ORF. Specifically, the Rasch model enhances our understanding of student ability by incorporating test item difficulty and providing a unified measure of “foundational reading skill” that encompasses both comprehension and precursor skills.

In contrast, GLMM, while computationally efficient, lacks the ability to capture differences between test items, making it less suitable for benchmark setting in educational assessments. However, the study also finds that IRT-derived CIs can be wider than what is considered technically satisfactory for reading science application, despite being acceptable from a psychometric standpoint.

When comparing different benchmarking approaches, it is essential to consider the tradeoffs in information utilization:

- Traditional benchmark setting relies on subject matter experts, providing precise but data-independent benchmarks.
- GLMM benchmarks leverage data and sample performance for high precision, but do not account for individual item difficulty.

- IRT-based benchmarks balance data context, sample performance and item-level performance, making it the most comprehensive approach despite potentially lower technical precision.

Thus, while IRT may be less precise than traditional methods, it makes the best use of available data and provides a more accurate representation of student ability.


Another major contribution of this study is the classical analysis of assessment unidimensionality and the integrity of literacy skill components within the broader latent construct of “foundational reading skill.” The study modeled precursor skills with RC, and the findings suggest that some skills—such as accuracy and ORF in connected text—serve as stronger predictors of RC than others, such as LS or LC. Even so, the set of skills measured by these assessments provide a great deal of reliability and unidimensionality. Additionally, some of the item differences appear to be sample- or curriculum-dependent, indicating the need for consideration as reading science experts deliberate further refinement of MPLs and the GPF using empirical psychometric analysis not previously available.

Finally, this study represents the largest compilation of foundational reading data available, with one possible exception being Crawford et al. (2024). However, unlike Crawford’s work, which focuses solely on reading science, this study conducted psychometric analyses on the data—assessments in 32 languages across eight countries. The study revealed language-group patterns that become apparent only when analyzing a large data set. While the study models just four language groups, this approach can be expanded as additional data becomes available. In parallel, UIS is conducting linguistic and reading sciences analyses (and later math science) to establish non-psychometric aspects of language group analysis.

Suitability of IRT for Benchmark Setting

IRT’s is particularly useful in benchmark setting because it models the probabilistic relationship between student ability and item difficulty, creating a continuous, interpretable proficiency scale. This approach ensures that performance benchmarks are defined in relation to an empirically derived latent ability continuum. The Rasch model effectively maps student responses along a common metric, facilitating meaningful comparisons across different assessments and test forms.

A key advantage of IRT-based benchmark setting is its ability to maintain measurement invariance. Unlike traditional norm-referenced methods, which rely on fixed percentiles and may be biased by sample characteristics, IRT ensures that proficiency levels are anchored to the underlying construct of reading proficiency. The study’s approach to defining precursor skill thresholds relative to RC benchmarks further underscores the utility of IRT in setting valid and consistent benchmarks across diverse contexts and across both comprehension and its precursor skills, and along the same latent scale. Additionally, CIs in IRT-derived estimates quantify the degree of uncertainty associated with proficiency classifications.



Furthermore, the ability of IRT to estimate student ability independently of the specific items administered makes it especially suitable for large-scale educational assessments. The ability to equate scores across different test forms enhances comparability, ensuring that the established benchmarks remain valid even when different test versions are used. This property is particularly advantageous in longitudinal studies and cross-sectional comparisons where maintaining consistency in measurement is essential.

Comparative Analysis: IRT vs GLMM

While GLMM is computationally efficient and provides precise estimates, it falls short in capturing item-level differences. By treating all items equally in the estimation of ability, it fails to account for important variations in item difficulty and discrimination. IRT-based models explicitly incorporate item-level parameters, allowing more accurate estimation of student proficiency across all skills.


Comparative analysis between RA-IRT and RA-GLMM results showed that while GLMM produces narrower confidence intervals, this precision comes at cost of ignoring variability in student responses. This trade-off highlights a fundamental issue in benchmark estimation—precision alone is not sufficient, if it comes at the expense of capturing the true distribution of ability levels. IRT embraces the complexity of item interactions and variability, making it a more reliable approach for large-scale educational assessments.

Limitations and Future Research

Despite its advantages, IRT-based benchmark setting presents certain challenges.

- **Broad CIs** – Initial non-RC subtask benchmarks showed wide CIs, reducing precision. This was mitigated through subtask-specific SEs, but further refinements are needed. These refinements, forthcoming in future drafts of this report, will rely on a more precise understanding of the nature of the assessments being benchmarked using a stronger dialogue or discussion between psychometricians and reading science experts.
- **Assumption of unidimensionality** – The Rasch model may not fully capture the complexity of literacy development, particularly in assessments that integrate multiple skill domains. Future research should explore the effect of other IRT models such as the 2PL and 3PL models and/or multidimensional IRT models in benchmark estimation to better account for interactions between different literacy components.
- **Small sample sizes** – Integration of Bayesian estimation methods to estimate IRT model parameters could provide more flexible priors, allowing for accurate benchmark estimation while maintaining the interpretability and theoretical advantages of IRT.

In conclusion, while both IRT and GLMM offer valuable insights for foundational literacy assessment benchmarks, IRT provides a more comprehensive approach by incorporating item differences and making the best use of available data. Future methodological advancements





should focus on further refining CI estimation, examining the effects of additional IRT models on benchmark estimation, comparing additional methods of parameter estimation, and enhancing the accuracy and applicability of literacy benchmarks by using data-driven methods.

DRAFT

References

- Baker, F.B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques, Second edition (2nd ed.)*. CRC Press. <https://doi.org/10.1201/9781482276725>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical Theories of Mental Test Scores*, Reading MA: Addison-Wesley.
- Chalmers, R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
- Crocker, L. and Algina, J. (1986) *Introduction to classical and modern test theory*. Harcourt, New York, 527.
- de Ayala, R.J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists (1st ed.)*. Psychology Press. <https://doi.org/10.4324/9781410605269>
- Everitt, B.S. and Dunn, G. (2001). *Applied multivariate data analysis*. 2nd Edition, Hodder Arnold, London. <http://dx.doi.org/10.1002/9781118887486>
- Fuchs, L.S., Fuchs, D., Hosp, M.K. & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5(3), 239-256. https://doi.org/10.1207/S1532799XSSR0503_3
- Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, 34(1), 111-117. <https://doi.org/10.1177/001316447403400115>
- Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. T. (2020). Estimating model-based oral reading fluency: A Bayesian approach. *Educational and Psychological Measurement*, 80(5), 847–869. <https://doi.org/10.1177/0013164419900208>
- Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2004) *Applied linear regression models*. 4th Edition, McGraw-Hill/Irwin, New York.



Lord, F.M. (1980). *Applications of item response theory to practical testing problems (1st ed.)*. Routledge. <https://doi.org/10.4324/9780203056615>

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Technical Report DAMTP 2009/NA06*, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 2009.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement*, 1(2), 233–247. <https://doi.org/10.1177/014662167700100209>

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.

About the American Institutes for Research®

Established in 1946, the American Institutes for Research® (AIR®) is a nonpartisan, not-for-profit institution that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of education, health, and the workforce. AIR's work is driven by its mission to generate and use rigorous evidence that contributes to a better, more equitable world. With headquarters in Arlington, Virginia, AIR has offices across the U.S. and abroad. For more information, visit [AIR.ORG](https://www.air.org).



AIR® Headquarters

1400 Crystal Drive, 10th Floor
Arlington, VA 22202-3289
+1.202.403.5000 | [AIR.ORG](https://www.air.org)

Notice of Trademark: "American Institutes for Research" and "AIR" are registered trademarks. All other brand, product, or company names are trademarks or registered trademarks of their respective owners.

Copyright © 2023 American Institutes for Research®. All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, website display, or other electronic or mechanical methods, without the prior written permission of the American Institutes for Research. For permission requests, please use the Contact Us form on [AIR.ORG](https://www.air.org).